

LAGRANGE MULTIPLIERS IN STOCHASTIC PROGRAMMING*

SJUR D. FLÅM†

Abstract. A Fritz John multiplier rule is given for discrete time, finite horizon, stochastic programs. Particular emphasis is placed on constraint qualifications that allow for the application of this rule in normal Lagrange form.

Key words. nonsmooth analysis, constraint qualifications, Kuhn-Tucker conditions

AMS(MOS) subject classification. 49B60

1. Introduction. Finite horizon, stochastic programs typically come in the following general form:

$$(P1) \quad \text{minimize the overall expected cost } \mathbb{E}f(\omega, x_1(\omega), \dots, x_T(\omega))$$

by making, sequentially, at each stage $t = 1, 2, \dots, T$, a decision $x_t(\omega) \in \mathbb{R}^{n_t}$, under imperfect information about the exact state ω of the world. Formally,

$$(1.1) \quad x_t(\cdot) \text{ should be } \Sigma_t\text{-measurable,}$$

where Σ_t , $t = 1, 2, \dots, T$, are known sub-sigma-algebras of a given probability space $(\Omega, \Sigma, \text{Pr})$. Most often, in practice, knowledge about the generic outcome ω in the sample space Ω becomes more refined with the passage of time, or, at least, does not deteriorate. To model such step-wise reduction of uncertainty we naturally assume that

$$(1.2) \quad \Sigma_1 \subset \Sigma_2 \subset \dots \subset \Sigma_T.$$

By way of example, let the information flow be generated by a stochastic process ξ_1, \dots, ξ_T on Ω . Then x_t should depend only on the actual realization of ξ_1, \dots, ξ_t ; i.e., Σ_t is, in this case, the smallest σ -algebra rendering all (possibly vector) variates ξ_1, \dots, ξ_t measurable, and (1.2) holds.

In addition to imposing the informational constraints (1.1) and (1.2), we also require that

$$(1.3) \quad g_t(\omega, x_1(\omega), \dots, x_t(\omega)) \leq 0 \quad \text{almost surely (a.s.) for } t = 1, \dots, T,$$

where $g_t(\cdot, x_1(\cdot), \dots, x_t(\cdot))$ is Σ_t -measurable with values in \mathbb{R}^{m_t} , and inequality (1.3) is understood to hold component-wise.

This completes the heuristic description of the multistage stochastic optimization problem. More formal assumptions are relegated to § 2.

The purpose of this paper is to characterize locally optimal solutions to problem (P1) in terms of the so-called *Lagrange multipliers* or Kuhn-Tucker conditions. It is well known, at least in deterministic programming, that such multipliers serve several ends expediently: They are prominent in optimality conditions, dominate much of stability analysis, play major roles in methods involving duality or decomposition, and are, not to forget, keys to the design of exact penalty functions. Thus, there is ample

* Received by the editors September 12, 1988; accepted for publication (in revised form) September 17, 1990. This paper was written at the University of Bayreuth. This research was partially supported by Ruhrgas and NAVF.

† Institute of Economics, University of Bergen, 5008 Bergen, Norway.

motivation for exploring their existence and nature in the context of stochastic programming. However, stochasticity present, the multipliers take on new features; they become random vectors, or they are measures. This has all been clearly brought to the fore by Rockafellar and Wets in a series of papers; see [17]–[19] and the references therein. Much of their analysis is confined to the case when $f(\omega, \cdot)$ and each component of $g_t(\omega, \cdot)$ is convex, with special attention paid to conditions that ensure existence of primal-dual optimal pairs of solutions. Their approach is the min-max theory of Lagrangian saddle-functions, such functions arising naturally from Rockafellar's scheme of conjugate duality and optimization of perturbed criteria [13].

By contrast, we will dispense with convexity assumptions and take existence of a locally optimal solution to (P1) for granted. Our main concern is to identify the precise nature of multipliers, to guarantee their existence, and finally, to provide constraint qualifications that imply that the multiplier rule has the desired normal form.

Our vehicle is the nonsmooth calculus of Clarke [2]. Hiriart-Urruty [7] has already applied this apparatus to 2-stage versions of our problem. He preferred to treat (P1) in its dynamic form and did not explore to what function space the second stage multipliers belong. As a difference, we avoid dynamic programming and instead analyze (P1) in its extended “static” form. The merit of this approach is that a multiplier rule emerges that completely parallels familiar results for deterministic programs.

To invoke the toolkit of nonsmooth calculus we need that *strategies* $x_t(\cdot)$ and *constraint functions* $g_t(\cdot, x_1(\cdot), \dots, x_t(\cdot))$, $t = 1, \dots, T$, belong to specified Banach spaces of measurable functions on $(\Omega, \Sigma, \text{Pr})$. In this regard we choose to steer away from two notable problem versions; namely,

- (i) when $g_t(\cdot, x_1(\cdot), \dots, x_t(\cdot))$, $t = 1, \dots, T$, are essentially bounded, and
- (ii) the case when strategies x_t , and constraints g_t are continuous with (1.3) sharpened to hold for all ω .

The reason for avoiding (i) and (ii) here is threefold. First, we plan to complement the studies of Rockafellar and Wets [17]–[19] by focusing on spaces of perturbations that have more pleasant duality properties than $L^\infty(\Omega)$ and $C(\Omega)$. Thus, [4] and [20] may be regarded as predecessors of this paper. Second, it may often be difficult, or even unnatural, to argue a priori that all g_t (and possibly also x_t) are essentially bounded or continuous on Ω . Frequently, larger spaces such as $L^p(\Omega)$, $1 \leq p < \infty$, are more convenient and realistic. Third, since $L^\infty(\Omega)$ and $C(\Omega)$ are “too small,” we are left with “too many” multipliers. Not only are many of these hard to interpret, but computations may also become very demanding. To illustrate this, consider a closed ball in the continuous dual $L^\infty(\Omega)^*$, large enough to contain all optimal multipliers and relevant approximations of these. Unless the sample space (Ω, Σ) is finite, this ball is neither w^* -metrizable nor w^* -sequentially compact. Then maximizing sequences of multipliers can, in principle, escape our control. The essential difficulty here stems from the fact that $L^\infty(\Omega)$ is either finite-dimensional (trivial case) or nonseparable (interesting case). By contrast, as we will see, all spaces $L^p(\Omega)$, $1 \leq p < \infty$, and some generalizations of these, are very well behaved. Thus, this paper serves both as a warning against the popular, quite common choice $L^\infty(\Omega)$, and as an invitation to consider alternative, more tractable spaces of perturbations.

In all events, we acknowledge that problem (P1) is very challenging when it comes to efficient computation. Even linear, 2-stage versions can often be solved only approximately; see [1] and [10]. Computational concerns are, however, beyond the scope of this paper, which is organized as follows. Section 2 provides some necessary technical prerequisites. Section 3 gives the multiplier rule, and the paper is concluded by briefly addressing the effect of standard constraint qualifications.

2. Preliminaries. This section collects all assumptions that are imposed to make problem (P1) well defined.

2.1. On probability spaces. We require that all σ -algebras $\Sigma_1, \dots, \Sigma_T$, mentioned in (1.1) and (1.2), be *complete*. If some Σ_t is not, it should be completed in the standard way: Include in Σ_t all subsets of Σ_t -measurable zero-sets. This done, we identify measurable functions (or events) that coincide almost surely. We also demand that $\Sigma_1, \dots, \Sigma_T$ be *separable* in the sense that each Σ_t becomes a separable metric space when endowed with the distance function

$$d(A, B) := \Pr(A \setminus B) + \Pr(B \setminus A).$$

2.2. On strategy spaces. First, we require that the decision x_t at stage t belongs to a space X_t of \mathbb{R}^{n_t} -valued, Σ_t -measurable random vectors. Thus, the restriction that strategies should be *nonanticipative* is implicit in our problem formulation. In addition, we insist that X_t be Banach. This amounts possibly to a stronger condition than (1.1). Also, we add here a constraint

$$(2.1) \quad x \in C,$$

where the set $C \subset X_1 \times \dots \times X_T$ is closed and nonempty. The abstract constraint (2.1) typically accounts for fairly tractable restrictions, for instance, upper and lower bounds, that may be imposed in addition to (1.2). We will give no further details on the precise nature of the set C .

2.3. On spaces of perturbations and their duals. The standard (and canonical) way to perturb problem (P1) is to replace (1.3) by

$$(2.2) \quad g_t(\omega, x_1(\omega), \dots, x_t(\omega)) \leq u_t(\omega) \quad \text{a.s. for } t = 1, \dots, T,$$

where the perturbation $u_t(\omega) \in \mathbb{R}^{m_t}$ belongs to some subspace U_t of measurable random vectors. We now proceed to identify an appropriate class of such function spaces U_t . In general, when Σ is a σ -algebra on Ω , denote by $L^o(\Sigma)$ the linear space of all Σ -measurable, real-valued functions on Ω . A linear subspace $\text{OCBF} \subset L^o(\Sigma)$ is said to be an *order continuous Banach foundation space* if there exists a norm $\|\cdot\|$ (making OCBF complete) such that

- (a) $\Theta \in L^o(\Sigma)$, $\Theta' \in \text{OCBF}$, $|\Theta(\omega)| \leq |\Theta'(\omega)|$ almost surely implies $\Theta \in \text{OCBF}$ and $\|\Theta\| \leq \|\Theta'\|$;
- (b) $\Theta'' \geq \Theta'^{n+1}$, $\Theta''(\omega) \downarrow 0$ almost surely and $\Theta'' \in \text{OCBF}$ imply $\lim_{n \rightarrow \infty} \|\Theta''\| = 0$;
- (c) The essential support of the entire space OCBF equals Ω .

Examples include the $L^p(\Omega)$ spaces, $1 \leq p < +\infty$, and their generalization to Orlicz spaces where the Δ_2 -condition holds [6], [10]. Also, when Ω is countable, the spaces ℓ^p , $1 \leq p < +\infty$, and c_o are of this type. Important properties of OCBF-spaces here are the following: *convergence in norm implies convergence in probability*, and *continuous linear functionals are representable as integrands*; that is, for the continuous dual OCBF^* we have

$$\text{OCBF}^* = \{\Theta^* \in L^o(\Sigma): E|\Theta \cdot \Theta^*| < +\infty \text{ for all } \Theta \in \text{OCBF}\}$$

with the action of $\Theta^* \in \text{OCBF}^*$ being defined by $\langle \Theta, \Theta^* \rangle := E\Theta \cdot \Theta^*$. For details see [10]. In the following we will speak of OCBF-spaces whose elements are random vectors in \mathbb{R}^m , say. On such occasions we tacitly assume that the OCBF-space in question arises as the product of m identical OCBF-spaces of real-valued functions. We assume that *the perturbation u_t in (2.2) belongs to an OCBF-space U_t for each $t = 1, \dots, T$* . Note that since $L^\infty(\Omega)$ is not an OCBF-space, this choice is excluded.

2.4. On the constraint functions g_t . A crucial assumption is that each g_t must map $X_1 \times \cdots \times X_t$ into an OCBF-space U_t of \mathbb{R}^{m_t} -valued, Σ_t -measurable random vectors, i.e.,

$$x_1 \in X_1, \dots, x_t \in X_t \Rightarrow g_t(\cdot, x_1(\cdot), \dots, x_t(\cdot)) \in U_t \quad \text{for } t = 1, \dots, T.$$

As an example, suppose $g_t(\omega, z)$ is affinely bounded, i.e.,

$$\|g_t(\omega, z)\| \leq |a_t(\omega) \cdot z + \beta_t(\omega)| \quad \text{a.s.}$$

with $a_t \in L^p(\Omega)$, $X_1 \times \cdots \times X_t \subset L^q(\Omega)$, $1 < p < +\infty$, $1/p + 1/q = 1$, and $\beta_t \in L^1(\Omega)$. Then $U_t = L^1(\Omega)$ would be a suitable choice. We remark that for the measurability of $g_t(\cdot, x_1(\cdot), \dots, x_t(\cdot))$ it suffices that g_t is Σ_t -normal [14].

2.5. On Lipschitz continuity. We assume that the criterion $\text{Ef}(\omega, x(\omega))$ is locally Lipschitz on an open set that contains C . For example, if X is a closed subspace of $L^\infty(\Omega)$ and

$$|f(\omega, x) - f(\omega, x')| \leq k(\omega)|x - x'|$$

for every pair of constant vectors x, x' with $k \in L^1(\Omega)$, then the above assumption holds with Lipschitz constant Ek , see [7, Lemma 3]. We also assume that *each*

$$g_t: X_1 \times \cdots \times X_t \rightarrow U_t, \quad t = 1, \dots, T,$$

is locally Lipschitz continuous on an open set containing C .

3. The multiplier rule. In keeping with the duality scheme of Rockafellar [13], suppose the perturbation $u_t \in U_t$ in (2.2) becomes available only at an extra expense

$$\langle y_t, u_t \rangle := E(y_t \cdot u_t).$$

Here y_t belongs to the dual space $Y_t := U_t^*$ (reflexiveness is not assumed, e.g., $U_t = L^1(\Omega)$, $Y_t = L^\infty(\Omega)$), and y_t should be interpreted as a random, exogenous price regime revealed only at stage t . It is appropriate to contemplate now what kind of measurability u_t and y_t could reasonably enjoy. For this purpose recall that our knowledge about the identity of ω is given, at stage t , only up to Σ_t . That is, we can discern for any event in Σ_t , and *only* such events, whether it has actually happened or not. Also, by assumption, each $g_t(\cdot, x_1(\cdot), \dots, x_t(\cdot))$ is Σ_t -measurable. Thus, when evaluating various perturbations u_t , it is impossible to discriminate between different candidates with finer precision than allowed for by Σ_t . This amounts to having u_t measurable with respect to Σ_t . The price y_t , unveiled in period t and prior to the choice of x_t , should, for the same reason, also be Σ_t -measurable. In summary, the opportunity to procure ourselves with perturbation profiles

$$u := (u_1, \dots, u_T) \in U := U_1 \times \cdots \times U_T$$

at the additional cost $\langle y, u \rangle = \langle y_1, u_1 \rangle + \cdots + \langle y_T, u_T \rangle$, with $y := (y_1, \dots, y_T)$ and $y_t \in Y_t = U_t^*$, leads naturally to the *Lagrangian*

$$L(x, y) := \inf \{ \text{Ef}(x) + \langle y, u \rangle \mid g_t(x_1, \dots, x_t) \leq u_t \in U_t \text{ a.s., } t = 1, \dots, T \}.$$

Trivially, $L(x, y) = -\infty$ if for some t we have $y_t < 0$ on a set of positive measure. Therefore, only nonnegative prices $y_t \geq 0$ almost surely are worthy of further consideration, and then the Lagrangian takes on the familiar form

$$L(x, y) = \text{Ef}(x) + \langle y, g(x) \rangle,$$

where

$$\langle y, g(x) \rangle := \sum_{t=1}^T \langle y_t, g_t(x_1, \dots, x_t) \rangle.$$

The original (primal) problem (P1) can be compactly restated as

$$(P2) \quad \inf_{x \in C} \sup_{y \in Y_+} L(x, y),$$

where Y_+ denotes the nonnegative cone of $Y = Y_1 \times \dots \times Y_T$. Its value $\inf (P2)$ majorizes that of the associate dual problem

$$(D) \quad \sup_{y \in Y_+} \inf_{x \in C} L(x, y).$$

A saddle point (x, y) of L would solve (P2) and (D) optimally with equal values and must necessarily satisfy the *Kuhn–Tucker conditions*:

$$(3.1) \quad 0 = \langle y, g(x) \rangle,$$

$$(3.2) \quad 0 \in \partial_x [L(x, y) + \delta_C(x)],$$

where ∂_x denotes the partial subdifferential of convex analysis and δ_C is the extended indicator of C . y is then called a *Lagrange multiplier* at x . However, when problem (P2) is nonconvex, it is often unrealistic to search for global saddle points. Instead we may have to contend, at best, with local versions of such points. Yet, as we plan to show, it is not wishful thinking to maintain (3.1) and (3.2) with ∂_x signifying the Clarke subdifferential [2]. Indeed, the following result offers close to complete evidence that *local* solutions to (P2) are supported by Lagrange multipliers. For the statement we need the following definition.

DEFINITION. The correspondence $(x, y) \rightarrow \partial_x \langle y, g(x) \rangle$ is said to be *closed* if for every sequence $(x^k, y^k, \nabla^k) \in X \times Y \times \partial_x \langle y^k, g(x^k) \rangle$ where $x^k \rightarrow x$ in norm, and $y^k \rightarrow y$, $\nabla^k \rightarrow \nabla$ in the w^* -topologies, we have $\nabla \in \partial_x \langle y, g(x) \rangle$.

Note in particular that this holds if g is C^1 .

THEOREM 3.1 (Fritz John rule in stochastic programming). *Under the hypothesis of § 2, let x be a locally optimal solution to problem (P2). Suppose that the correspondence $(x, y) \rightarrow \partial_x \langle y, g(x) \rangle$ is closed. Then for every $\delta, \eta > 0$ sufficiently small there exists a nonzero, nonnegative multiplier*

$$(y_0, y_1, \dots, y_T) \in R \times Y_1 \times \dots \times Y_T$$

such that

$$(3.3) \quad y_t(\omega) \cdot g_t(\omega, x_1(\omega), \dots, x_t(\omega)) = 0 \quad \text{a.s.} \quad \text{for } t = 1, \dots, T,$$

and

$$(3.4) \quad 0 \in \partial_x \left[y_0 \operatorname{Ef}(x) + \sum_{t=1}^T \langle y_t, g_t(x_1, \dots, x_t) \rangle + \eta^{-1} \rho(y) d(x) \right],$$

where d denotes the distance function to

$$C_\delta(x) := \{x' \in C : \|x' - x\| \leq \delta\},$$

and ρ is a strict norm on the dual of $R \times U$ such that the ρ -topology on bounded sets coincides with the w^* -topology.

Remark. The role of the parameter δ is to restrict attention to a part $C_\delta(x)$ of C within which x is globally optimal. The other parameter η should be chosen so small that η^{-1} exceeds the Lipschitz constant of the function

$$[\text{Ef}(\cdot, x(\cdot)), g_1(\cdot, x_1(\cdot)), \dots, g_T(\cdot, x_1(\cdot)), \dots, x_T(\cdot)]$$

near x .

The proof of Theorem 3.1 will be simplified if we isolate and divorce some central arguments from their specific context of the theorem. Therefore, we will focus first on the generalized program

$$(P3) \quad \text{minimize } f_0(x) \text{ subject to } f_1(x) \leq 0 \quad \text{and} \quad x \in C,$$

and argue later that problem (P2) is only a special instance of (P3). In (P3) the set C is again a nonempty closed part of some Banach space X , and $f_0: X \rightarrow \mathbb{R}$, $f_1: X \rightarrow U$ are both locally Lipschitz near any point in C . The space U (of perturbations) is normed *separable* (not necessarily complete) and ordered by a relation \leq , defined, as usual, via a closed convex cone U_+ having

$$U_+^* := \{u^* \in U^*: \inf \langle u^*, U_+ \rangle = 0\}$$

as positive dual (polar cone).

We say that a *local solution* x to the generalized program (P3) is *supported by a multiplier* $(y_0, y_1) \in \mathbb{R}_+ \times U_+^*$ if $(y_0, y_1) \neq 0$, $\langle y_1, f_1(x) \rangle = 0$, and

$$(3.5) \quad 0 \in \partial_x [y_0 f_0(x) + \langle y_1, f_1(x) \rangle + \eta^{-1} \rho(y) d(x)],$$

where d is the distance function to

$$C_\delta(x) := \{x' \in C: \|x' - x\| \leq \delta\},$$

$\delta > 0$ being so small that x becomes a global solution to (P3) when restricted to $C_\delta(x)$ instead of C , and η^{-1} is greater than the Lipschitz constant of (f_0, f_1) near x . In (3.5) ρ is, as in (3.4), a strict norm on the dual of $R \times U$ generating the w^* -topology on bounded sets.

THEOREM 3.2 (Fritz John rule). *Under the above hypothesis on (P3) suppose that the correspondence $(x, y_1) \rightarrow \partial_x \langle y_1, f_1(x) \rangle$ is closed. Then every local solution to (P3) is supported by a multiplier.*

Proof. The statement is close to a transcription of Clarke's multiplier rule [2]. He deals, however, with only finitely many explicit restrictions (i.e., U is finite-dimensional). We must therefore carefully arrange the situation so that his method of proof carries over.

The main object is the pointwise maximum function $F: X \rightarrow \mathbb{R}$ defined, for arbitrary fixed parameter $\varepsilon > 0$, by

$$(3.6) \quad F(x') := \max_{y \in M} \langle y, (f_0(x') - f_0(x) + \varepsilon, f_1(x')) \rangle,$$

where $M \subset \mathbb{R} \times U_+^*$ is an appropriate set of multipliers. Specifically, since U is normed separable, there exists, by the Clarkson–Rieffel renorming lemma [9], a *strict* norm ρ_1 on the continuous dual space U^* , which is weaker than the usual dual norm, and such that the ρ_1 -topology on any bounded subset of U^* coincides with the w^* -topology. Now define a strict norm ρ on $\mathbb{R} \times U^*$ by

$$\rho(y_0, y_1) := (y_0^2 + \rho_1(y_1)^2)^{1/2},$$

and let

$$M := \{y \in \mathbb{R} \times U_+^*: \rho(y) \leq 1\}.$$

Note that $y \in M$ implies $\|y\| \leq \beta$ for some uniform bound β . We claim that F , as defined in (3.6), is Lipschitz near any $x' \in C$. Indeed, F being the pointwise maximum of the function family $\langle y, \varphi(\cdot) \rangle$, $y \in M$, with

$$(3.7) \quad \varphi(x') := (f_0(x') - f_0(x) + \varepsilon, f_1(x')),$$

it suffices to show that this family is equi-Lipschitz. That property follows readily from

$$\begin{aligned} |\langle y, \varphi(x') \rangle - \langle y, \varphi(x'') \rangle| &\leq \|y\| \|\varphi(x') - \varphi(x'')\| \\ &= \beta \|\varphi(x') - \varphi(x'')\|, \end{aligned}$$

and the Lipschitz continuity of φ (3.7).

We claim that $F(x') > 0$ for all $x' \in C$ sufficiently near x . In fact, $F(x') \leq 0$ would imply, in the first place, that $f_1(x') \leq 0$, because if $-f_1(x') \notin U_+$, then for some $u^* \in U_+^*$ we have $\langle u^*, -f_1(x') \rangle < 0$, and this contradicts $F(x') \leq 0$. Second, to posit the three conditions: $F(x') \leq 0$, $x' \in C$ is near x , and $f_1(x') \leq 0$, would entail the following contradiction:

$$f_0(x') \leq f_0(x) - \varepsilon.$$

Since x is locally ε -optimal, there exists by Ekeland's variational principle [2], a point x_ε , within $\sqrt{\varepsilon}$ -distance from x , that minimizes

$$F(x') + \sqrt{\varepsilon} \|x' - x_\varepsilon\|$$

over $C_\delta(x)$. Now the distance (function) d to $C_\delta(x)$ can be used for the purpose of exact penalization [2, Prop. 2.4.3]: To wit, x_ε is an *unconstrained* minimum for the function

$$F(x') + \sqrt{\varepsilon} \|x' - x_\varepsilon\| + \eta^{-1} d(x').$$

Therefore, letting B^* denote the closed unit ball in X^* , we have

$$(3.8) \quad 0 \in \partial G(x_\varepsilon) + \sqrt{\varepsilon} B^*,$$

where, using (3.7) and the fact that F is positive near x ,

$$(3.9) \quad G(x') := \max_{y \in M} [\langle y, \varphi(x') \rangle + \eta^{-1} \rho(y) d(x')].$$

We next intend to employ Theorem 2.8.2 of Clarke [2] for the estimation of $\partial G(x')$. For this, observe that the spaces $\mathbb{R} \times U$ and $\mathbb{R}^* \times U^*$ are placed in duality by the natural pairing

$$\langle (r, u), (r^*, u^*) \rangle := rr^* + \langle u, u^* \rangle.$$

Endow M , under this pairing, with the (relative) w^* -topology. Then M , being closed bounded, is compact by the Alaoglu–Bourbaki theorem. Moreover, M is metrizable because U is separable [9]. In particular,

(i) M is sequentially compact.

In addition we observe that

- (ii) the map $\langle y, \varphi(x') \rangle + \eta^{-1} \rho(y) d(x')$, which occurs in (3.9), is w^* -continuous in y on M for every fixed x' ;
- (iii) each function $\langle y, \varphi(x') \rangle + \eta^{-1} \rho(y) d(x')$, $y \in M$, in (3.9) is locally Lipschitz in x' , and the set

$$\{\langle y, \varphi(x') \rangle + \eta^{-1} \rho(y) d(x') : y \in M\}$$

is bounded.

The upshot of (i)–(iii) and the metrizable of M is that Theorem 2.8.2 of Clarke [2] is indeed applicable for the estimation of $\partial G(x_\varepsilon)$. Since $F(x_\varepsilon)$ and $G(x_\varepsilon)$ are positive, any maximizing multiplier $y_\varepsilon \in M$ in (3.6), as well as in (3.9), for $x' = x_\varepsilon$, must satisfy $\rho(y_\varepsilon) = 1$. Moreover, since ρ is strict this y_ε is unique. It follows from (3.8) and (3.9) and the closedness of $(x, y_1) \rightarrow \partial_x \langle y_1, f_1(x) \rangle$ that

$$(3.10) \quad 0 \in \partial_x [\langle y_\varepsilon, \varphi(x_\varepsilon) \rangle + \eta^{-1} d(x_\varepsilon)].$$

Now the arguments of Clarke [2, Thm. 6.1.1] apply verbatim and we contend with a sketch: Let some sequence $\varepsilon(k)$, $k = 1, 2, \dots$, tend downward to 0. Then $x_{\varepsilon(k)} \rightarrow x$, and some subsequence $y_{\varepsilon(k)}$, $k \in K$, will w^* -converge to a point $y \in M$. Since $\rho(y) = 1$ we have $y \neq 0$. Letting $k \in K$ pass to infinity in (3.10) we obtain the desired conclusion. \square

Remarks. It is crucial in the above proof that closed bounded subsets of $Y = U^*$ be w^* -compact and metrizable. For this, it is both necessary and sufficient that the normed space U is separable [9]. In particular, a nonseparable space $U = L^\infty(\Omega)$ does not fit our framework.

Evidently, we can identify ρ with the usual dual norm whenever the latter is strictly convex and U is reflexive. In particular, the proof can be simplified when $U = L^p(\Omega)$, $p \in (1, \infty)$. We wish, however, to accomodate for the nonreflexive space $U = L^1(\Omega)$ as well as c_0 , and then the dual norm is unsuitable.

The case when U equals the space $C(\Omega)$ of continuous functions on a compact set Ω in an Euclidean space, falls under the hypothesis of Theorem 3.2. Thus, in particular, we may give multiplier rules for semi-infinite programming.

Proof of Theorem 3.1. Recall that perturbations u_i in (2.2) are to be selected from an OCBF-space U_i of Σ_i -measurable random vectors in \mathbb{R}^m . The associated convex cone U_{i+} is, of course, the set of all $u_i \in U_i$ such that $u_i \geq 0$ almost surely. Since Σ_i is separable, so is U_i (see [10]) as well as

$$U := U_1 \times \dots \times U_T.$$

The said convex cone U_{i+} is closed because convergence in norm implies convergence in probability. Hence

$$U_+ := U_{1+} \times \dots \times U_{T+}$$

is also closed. Now let $f_0(x) := \text{Ef}(\omega, x(\omega))$,

$$f_1(x) := (g_1(\cdot, x_1(\cdot)), \dots, x_t(\cdot))_{t \in \{1, \dots, T\}}$$

and appeal to Theorem 3.2. \square

Remarks. As in Clarke [2, Thm. 2.7.5] conditions can be given that, when $C = X$, ensure that (3.4) holds almost surely; that is,

$$0 \in \partial_x [y_0 f(\omega, x(\omega)) + \sum_{t=1}^T y_t(\omega) g_t(\omega, x_1(\omega), \dots, x_t(\omega))] \text{ a.s.}$$

By appropriately redefining the cones U_{i+} , $i = 1, \dots, T$, there is no problem in accomodating for equality constraints.

As mentioned, we could design a more direct proof especially adapted to the instance $U = L^p(\Omega)$, $p \in (1, \infty)$. However, the important spaces $L^1(\Omega)$ and c_0 would then call for separate discussion.

4. Constraint qualifications. For the purpose of duality theory, stability analysis, exact penalty methods [5], and the like we need to ensure that the only multiplier

$y = (y_0, y_1, \dots, y_T)$ satisfying (3.3) and (3.4) with $y_0 = 0$, is the origin 0. We then say that the *multiplier rule holds in normal Lagrange form*.

Conditions guaranteeing the existence of Lagrange multipliers have been extensively studied in the literature on programming in Banach spaces: With reference to (P3) [21] offers the following sufficient condition:

$$(4.1) \quad 0 \in \text{int}_U [f'_1(x) T_C x - T_{U_-} f_1(x)]$$

provided $C \subset X$ is closed convex, f_0 is Frechet differentiable, and f_1 is continuously Frechet differentiable. Here $T_C x$ and $T_{U_-} f_1(x)$ denote the usual tangent cone (of convex analysis) of C at x , and $U_- := -U_+$ at $f_1(x)$, respectively. In particular [12], (4.1) would hold if

$$(4.2) \quad x \in \text{int } C \text{ and } f'_1(x) \text{ is surjective,}$$

or if

$$(4.3) \quad \text{there exists } x' \in T_C x \text{ such that } f'_1(x)x' \in \text{int } T_{U_-} f_1(x).$$

These conditions are stringent; however, (4.2) requires that the constraint $x \in C$ is not binding and that $f'_1(x)$ has full rank; (4.3) implies that U_+ has nonempty interior whereas in our context $\text{int } L^p(\Omega)_+ = \emptyset$ for all $p \in [1, \infty)$. Also, it is desirable to avoid the strong differentiability assumption behind (4.1).

For our approach, recall that in smooth deterministic programming, with $C = X$, a minimal hypothesis for existence of Lagrange multipliers is the well-known Mangasarian-Fromowitz (M-F) constraint qualification. In fact, (4.1) is then equivalent to M-F. If, moreover, data are convex, the M-F-qualification reduces to the Slater condition. The M-F-condition has also been taken into the context of nondifferentiable programming [2], [8], [15], and [16]. Our purpose here is twofold: first, we show that the M-F and Slater conditions carry easily over to program (P3), second, that they come naturally in terms of each realization ω , rather than intervening at the level of the functional spaces.

We begin with the *M-F-constraint qualification*, which is most general. According to this there should exist a direction $d = (d_1, \dots, d_T) \in X$ in the *Clarke tangent cone* $T_C(x)$ of C at x such that $g_t(\omega, x_1(\omega), \dots, x_t(\omega)) = 0$ implies

$$g_t^0(\omega, x_1(\omega), \dots, x_t(\omega); d_1(\omega), \dots, d_t(\omega)) < 0 \quad \text{a.s. for } t = 1, \dots, T.$$

Here g_t^0 denotes, for each given ω , the *Clarke directional derivative* of $g_t(\omega, x_1(\omega), \dots, x_t(\omega))$ at the point $(x_1(\omega), \dots, x_t(\omega))$ in the direction $(d_1(\omega), \dots, d_t(\omega))$. We tacitly assume that $g_t(\omega, z)$ is Lipschitz in z almost surely.

PROPOSITION 4.1. *Under the hypothesis of Theorem 3.1 and the M-F-constraint qualification, all multipliers at the local solution x to P have normal form.*

Proof. Suppose not. Then there exists $(y_1, \dots, y_T) \geq 0$, and different from zero, such that (for $\delta, \eta > 0$ sufficiently small)

$$0 \in \partial_x \left[\sum_{t=1}^T \langle y_t, g_t(x) \rangle + \eta^{-1} \|y\| d(x) \right].$$

Consequently,

$$\begin{aligned} 0 &\leq \left[\sum_{t=1}^T \langle y_t, g_t(\cdot) \rangle + \eta^{-1} \|y\| d(\cdot) \right]^0(x; d) \\ &\leq \sum_{t=1}^T \langle y_t, g_t^0(x_1, \dots, x_t; d_1, \dots, d_t) \rangle < 0, \end{aligned}$$

a contradiction. Here, the penultimate inequality follows from two facts: First, the

directional derivative of a sum minorizes the sum of directional derivatives [2, Prop. 2.3.3], and second, the derivative of the distance function in a tangent direction equals zero [2, § 2.4]. \square

Next we discuss the Slater condition.

PROPOSITION 4.2. *Suppose there exists some $\hat{x} \in C$ such that*

$$(4.1) \quad g_t(\omega, \hat{x}_1(\omega), \dots, \hat{x}_t(\omega)) < 0 \quad \text{a.s. for } t = 1, \dots, T,$$

with C and g_1, \dots, g_T being starshaped with respect to \hat{x} . Then the M-F-constraint qualification holds.

Proof. Starshapedness of a set with respect to a point means that the set can be entirely illuminated by a source of light placed at the point in question. Correspondingly, a function is starshaped at a point if and only if its epigraph is starshaped at that point. Select the direction $d := \hat{x} - x$ and note that $g_t(\omega, x_1(\omega), \dots, x_t(\omega)) = 0$ implies, together with (4.1) and the star-shapedness that

$$g_t^0(\omega, x_1(\omega), \dots, x_t(\omega); d_1(\omega), \dots, d_t(\omega)) < 0 \quad \text{a.s.}$$

An appeal to Proposition 4.1 completes the proof. \square

REFERENCES

- [1] J. R. BIRGE AND S. W. WALLACE, *A separable piecewise linear upper bound for stochastic linear programs*, SIAM J. Control Optim., 26 (1988), pp. 725-739.
- [2] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.
- [3] J. DIESTEL, *Sequences and Series in Banach Spaces*, Springer-Verlag, Berlin, 1984.
- [4] M. EISNER AND P. OLSEN, *Duality for stochastic programming, interpreted as L. P. in L_p -space*, SIAM J. Appl. Math., 28 (1974), pp. 779-792.
- [5] S. D. FLÅM AND J. ZOWE, *Exact penalty functions in single-stage stochastic programming*, Optimization, to appear.
- [6] E. GINER, *Espace intégraux de type orlicz; Dualité, compacité, convergence en mesure et applications à l'optimisation*, Ph.D. thesis, Centre Universitaire de Perpignan, 1977.
- [7] J.-B. HIRIART-URRUTY, *Conditions nécessaires d'optimalité pour un programme stochastique avec recours*, SIAM J. Control Optim., 16 (1978), pp. 317-329.
- [8] ———, *Refinements of necessary optimality conditions in nondifferentiable programming I*, Appl. Math. Optim., 5 (1979), pp. 63-82.
- [9] R. B. HOLMES, *Geometric Functional Analysis and its Applications*, Springer-Verlag, Berlin, 1975.
- [10] P. KALL, *Stochastic programming with recourse. Upper bounds and moment problems*, Math. Res., 45 (1988), pp. 86-103.
- [11] L. V. KANTOROVICH AND G. P. AKILOV, *Functional Analysis*, Pergamon Press, Oxford, 1982.
- [12] S. KURCYUZ, *On the existence and nonexistence of Lagrange multipliers in Banach spaces*, J. Optim. Theory Appl., 20 (1976), pp. 81-110.
- [13] R. T. ROCKAFELLAR, *Conjugate Duality and Optimization*, CBMS-NSF Regional Conf. Ser. in Appl. Math., Vol. 16, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1974.
- [14] ———, *Integral functionals, normal integrands and measurable selections*, in Nonlinear Operators and Calculus of Variations, Lecture Notes in Math., Vol. 543, Springer-Verlag, New York, Berlin, 1977.
- [15] ———, *Extensions of subgradient calculus with applications to optimization*, Nonlinear Anal., 9 (1985), pp. 665-698.
- [16] ———, *Lagrange multipliers and subderivatives of optimal value functions in nonlinear programming*, Math. Programming Stud., 17 (1982), pp. 28-66.
- [17] R. T. ROCKAFELLAR AND R. J.-B. WETS, *Measures as Lagrange multipliers in multistage stochastic programming*, J. Math. Anal. Appl., 60 (1977), pp. 301-313.
- [18] ———, *The optimal recourse problem in discrete time*, SIAM J. Control Optim., 16 (1978), pp. 16-36.
- [19] ———, *Deterministic and stochastic optimization problems of Bolza type in discrete time*, Stochastics, 10 (1983), pp. 273-312.
- [20] R. J.-B. WETS, *Problèmes d'aux en programmation stochastique*, C. R. Acad. Sci. Paris, Ser. A-B, 270 (1970), pp. 47-50.
- [21] J. ZOWE AND S. KURCYUZ, *Regularity and stability of the mathematical programming problem in Banach spaces*, Appl. Math. Optim., 5 (1972), pp. 49-62.

ALGEBRAIC GEOMETRIC ASPECTS OF FEEDBACK STABILIZATION*

SHIVA SHANKAR† AND V. R. SULE‡

Abstract. This paper develops a theory of feedback stabilization for SISO transfer functions over a general integral domain which extends the well-known coprime factorization approach to stabilization. Necessary and sufficient conditions for stabilizability of a transfer function in this general setting are obtained. These conditions are then refined in the special cases of unique factorization domains (UFDs), Noetherian rings, and rings of fractions. It is shown that these conditions can be naturally interpreted geometrically in terms of the prime spectrum of the ring. This interpretation provides a natural generalization to the classical notions of the poles and zeros of a plant.

The set of transfer functions is topologized so as to restrict to the graph topology of Vidyasagar [*IEEE Trans. Automatic Control*, AC-29 (1984), pp. 403–418], when the ring is a Bezout domain. It is shown that stability of a feedback system is robust in this topology when the ring is a UFD.

This theory is then applied to the problem of stabilization of multidimensional systems. The above stabilizability criterion is interpreted geometrically in terms of affine varieties in \mathbb{C}^n when the stability region is the complement of a compact polynomially convex domain Γ . This criterion restricts to the well-known result for two-dimensional systems when Γ is the unit polydisc; it also allows the resolution of an open problem of Guiver [*Multidimensional Systems Theory*, D. Reidel, 1985]. Finally it is shown that while feedback stabilizability is robust, it is not, however, a generic property.

Key words. stabilization, graph topology, unique factorization domain (UFD)

AMS(MOS) subject classifications. 93D15, 93D25, 93B27, 93B25

1. Introduction. In this paper we develop a general theory for feedback stabilization of plants whose transfer functions are described by fractions over a general *integral domain*. This ring theoretic stabilization problem is patterned after Desoer et al. [4], Vidyasagar, Schnider, and Francis [10] and is defined as follows.

1.1. The stabilization problem. Let \mathbf{A} be an integral domain that represents the ring of stable causal SISO transfer functions. Denote by \mathbf{F} the field of fractions of \mathbf{A} consisting of all possible transfer functions. Define the subset \mathcal{F} of $\mathbf{F} \times \mathbf{F}$ as follows:

$$\mathcal{F} = \{(p, c) \in \mathbf{F} \times \mathbf{F} \mid 1 + pc \neq 0\}.$$

For a given transfer function p in \mathbf{F} the *fibre* over p , denoted \mathcal{F}_p , is the set of all transfer functions c for which (p, c) belongs to \mathcal{F} . Note that \mathcal{F}_p is never empty; indeed \mathcal{F}_0 equals \mathbf{F} , and \mathcal{F}_p for every nonzero p is the complement of a singleton in \mathbf{F} .

DEFINITION. The pair (p, c) in \mathcal{F} is said to be *stable* if $(1+pc)^{-1}$, $p(1+pc)^{-1}$, and $c(1+pc)^{-1}$ all belong to the subring \mathbf{A} of \mathbf{F} . In this case c is said to *stabilize* the transfer function p .

THE STABILIZATION PROBLEM. Given p in \mathbf{F} determine the set of all c in the fibre over p that stabilizes p .

Remark. The above “ring theoretic” stabilization problem specializes to the classical input/output stabilization of a dynamical system when \mathbf{A} is a ring of operators on the space of bounded functions (called inputs) to another space of bounded functions

* Received by the editors October 20, 1989; accepted for publication (in revised form) October 26, 1990.

† Department of Electrical Engineering, Indian Institute of Technology, Powai, Bombay 400076, India.

‡ Department of Electrical Engineering, Indian Institute of Technology, Kanpur 208016, India.

(called outputs); define the map:

$$(1) \quad \begin{aligned} H: \mathcal{F} &\rightarrow \mathbf{F}^{2 \times 2}, \\ (p, c) &\mapsto \begin{bmatrix} (1+pc)^{-1} & -p(1+pc)^{-1} \\ c(1+pc)^{-1} & (1+pc)^{-1} \end{bmatrix}. \end{aligned}$$

The map H represents the input/output map

$$[u_1, u_2]^T \mapsto [e_1, e_2]^T$$

of the *feedback system* $\langle p, c \rangle$ shown in Fig. 1 with p and c denoting the transfer functions of the *plant* and *controller*, respectively, where u_1, u_2 are bounded inputs.

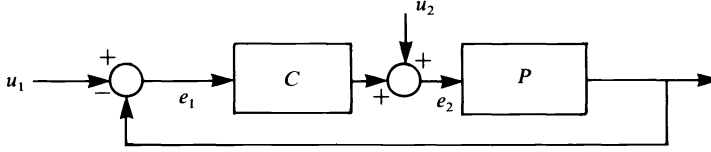


FIG. 1. Feedback system.

The feedback system $\langle p, c \rangle$ is said to be stable if the image $H(p, c)$, belongs to $\mathbf{A}^{2 \times 2}$, in which case c is called a *stabilizing controller* of p .

Note that for some p , the set of c in \mathcal{F}_p that stabilize it could be empty. This situation is more general than those considered in (Desoer et al. [4], Vidyasagar [9]) where the plant is always assumed to have coprime fractions, and because of which the set of stabilizing controllers is always nonempty.

1.2. Motivation. The above ring theoretic stabilization problem is solved in Desoer et al. [4] for those p admitting coprime fractions over the ring \mathbf{A} . Its subsequent extensions by Desoer and Gustavson [3] and Vidyasagar [9] also relied on the existence of coprime fractions for p . Moreover, a basis for the *graph topology* on the set of plants introduced by Vidyasagar [8] to study robustness of stability is defined again via coprime fractions. Thus coprime factorizability is central to the stabilization theory developed so far.

In many engineering situations, for instance multidimensional systems [2], spatially distributed systems (see Kamen [2]) etc., plant transfer functions do not always admit coprime fractions. Thus from these examples, as well as from a theoretical point of view, it is desirable to have a more general theory of stabilization which subsumes the existing theory whenever coprime fractions do indeed exist. It is also important to be able to interpret geometrically this general algebraic theory. For instance, in the case of plants whose transfer functions are fractions of functions holomorphic in some domain in \mathbb{C} , the stabilizability of the plant is determined geometrically in terms of its poles and zeros. This is the case with linear time invariant (LTI) lumped or distributed plants. Similarly, for two-dimensional (2-D) plants it is now well known (Bose [2]) that the stabilizability of the plant depends on whether the pole-zero sets of the plant intersect within the unit bidisc. What then are the geometric objects in general rings (which are the generalizations of the poles and zeros of the above special cases) in terms of which the stabilizability conditions of this more comprehensive theory could be expressed? It is also necessary to topologize the set of plants which will specialize to the graph topology of Vidyasagar in order to study robustness of stability. These considerations motivate the following questions:

(1) What are necessary and sufficient conditions for stabilizability of a plant with fractions over a general integral domain, (a question originally posed by Vidyasagar, Schnider, and Francis [10])? For instance, in the study of 2-D systems by Guiver and Bose [2] existence of rational coprime fractions is also necessary for stabilizability.

(2) How can the above conditions be interpreted in special cases such as Bezout domains, unique factorization domains (UFDs), Noetherian rings, rings of fractions, etc.?

(3) What are the geometric interpretations of the above algebraic criteria?

(4) What is a natural topology on the set of all plants? Is stabilizability robust with respect to this topology? Does the set of stabilizable plants form a dense set in the set of all plants?

(5) How can the above results be interpreted geometrically for the special case of the n -D stabilization problem?

In this paper we answer the above questions for SISO plants. Though we restrict ourselves to the SISO case, more for clarity and for the natural geometric interpretations that can be obtained, the results nevertheless provide a stabilizability theory for a diverse class of linear systems including n -D, distributed, and infinite-dimensional systems. We have recently obtained corresponding algebraic criteria for feedback stability in the MIMO case over integral domains. Its ramifications such as tracking, disturbance rejection, among others, will be reported elsewhere.

1.3. Outline of results. As the main body of the paper is somewhat technical we provide a detailed preview of the main results of this paper. This outline also serves to fix notation. For general background in commutative algebra we refer the reader to Atiyah and McDonald [1]; for algebraic geometry, refer to Hartshorne [6, Chap. 1].

1.3.1. General stabilizability conditions. In § 2.1 we present the following necessary and sufficient condition for stabilizability of a plant with fractions over a general integral domain. Let $p = nd^{-1}$ be a transfer function in \mathbf{F} . Define the following ideal quotients of the principal ideals (n) and (d) :

$$\begin{aligned}\mathbf{a} &= ((n): d) \quad \text{denoted hereafter as } (n: d), \\ \mathbf{b} &= (d: n).\end{aligned}$$

These ideals are independent of the fraction representing p .

One of our principal results is the following theorem (cf. Theorem 2.1.1).

THEOREM 2.1.1'. *The transfer function $p = nd^{-1}$ is stabilizable if and only if the ideals \mathbf{a} and \mathbf{b} are coprime (i.e., $\mathbf{a} + \mathbf{b} = \mathbf{A}$).*

We next consider the interpretation of this result for the following special cases.

(1) The ring \mathbf{A} is a Bezout domain: Here we show that \mathbf{a} and \mathbf{b} are always coprime; hence every transfer function is stabilizable (Corollary 2.1.4).

(2) When \mathbf{A} is a UFD the transfer function p is stabilizable if and only if p has coprime fractions (Corollary 2.1.5).

(3) When \mathbf{A} is a Noetherian ring the stabilizability condition can be directly expressed in terms of the primary decompositions of (n) and (d) (Corollary 2.1.7).

(4) Let \mathbf{A} be a ring of fractions $\mathbf{S}^{-1}\mathbf{B}$ of a ring \mathbf{B} with respect to a multiplicatively closed subset \mathbf{S} of \mathbf{B} . Then the above ideals \mathbf{a} and \mathbf{b} in \mathbf{A} are extensions of ideals \mathbf{I}_a and \mathbf{I}_b in \mathbf{B} , respectively. The stabilizability condition is now as in the following corollary (cf. Corollary 2.1.8).

COROLLARY. *The transfer function p is stabilizable if and only if*

$$(\mathbf{I}_a + \mathbf{I}_b) \cap \mathbf{S} \neq \emptyset.$$

1.3.2. Geometric interpretations. In § 2.2 we discuss the geometric implications of the above results. By “geometric” we mean interpretations in terms of prime ideals of the ring. In the well-known case of linear time invariant (lumped as well as distributed) plants, the set of stable transfer functions is a ring of functions holomorphic in some subset of \mathbb{C} . In such cases the zeros of a stable transfer function correspond to prime ideals of the ring containing the transfer function. Hence for a general ring \mathbf{A} the interpretation of the stabilizability condition in terms of prime ideals of \mathbf{A} is a natural generalization of the interpretation of stabilizability in terms of poles and zeros. This point is developed in Sule [7].

Let $\text{spec } \mathbf{A}$ denote the prime spectrum of the ring \mathbf{A} . For an ideal \mathbf{c} of \mathbf{A} let $V(\mathbf{c}) \subseteq \text{spec } \mathbf{A}$ denote the set of all prime ideals of \mathbf{A} containing \mathbf{c} . Clearly, Theorem 2.1.1 is equivalent to

$$V(\mathbf{a}) \cap V(\mathbf{b}) = \emptyset.$$

We show that when \mathbf{A} is a Noetherian ring $V(\mathbf{a})$ and $V(\mathbf{b})$ can be determined from the algebraic criteria obtained above. Thus in this case the stabilizability condition is interpreted directly in terms of n and d (Corollary 2.1.7) and hence, in light of the above remarks, directly in terms of notions which are generalizations of the poles and zeros of the plant.

We then consider the special case when $\mathbf{A} = \mathbf{S}^{-1}\mathbf{B}$. Define the subset $\Omega \subseteq \text{spec } \mathbf{B}$

$$\Omega = \{\mathbf{p} \in \text{spec } \mathbf{B} \mid \mathbf{p} \cap \mathbf{S} = \emptyset\}.$$

The geometric equivalent of Corollary 2.1.8 is the following.

COROLLARY 2.1.8'. *The transfer function p is stabilizable if and only if*

$$V(\mathbf{I}_a) \cap V(\mathbf{I}_b) \cap \Omega = \emptyset.$$

This is a key result in our development since it is this result that we interpret in the special case of n -D systems in terms of the geometry of affine varieties in \mathbb{C}^n .

1.3.3. Robustness results. In § 2.2 we topologize the set of all transfer functions with the purpose of studying genericity questions. Here we consider the field \mathbf{F} of transfer functions as a quotient space arising from an equivalence relation on $\mathbf{A} \times (\mathbf{A} \setminus \{0\}) := \mathbf{A} \times \mathbf{A}^*$. Hence the natural topology to consider is the quotient topology inherited from a topology (not necessarily the product topology) on $\mathbf{A} \times \mathbf{A}^*$. We show that in the special case of a Bezout domain this is just the well-known graph topology.

For a topological ring \mathbf{A} satisfying additional conditions we show that stability of a feedback system is a robust property in the quotient topology.

1.3.4. Multidimensional stabilization. In § 3 we apply the above theory to obtain necessary and sufficient conditions for stabilizability of n -D plants, as well as study the robustness of stabilizability. This problem belongs to one of the above special cases, namely when the ring \mathbf{A} is a ring of fractions $\mathbf{S}^{-1}\mathbf{B}$ where \mathbf{B} is the polynomial ring in n indeterminates over a subfield \mathbb{K} of \mathbb{C} , denoted $\mathbb{K}[X_1 \cdots X_n]$. We show here that the central issue in geometrically interpreting the stabilizability condition is that of characterizing affine varieties corresponding to prime ideals of \mathbf{B} that do not meet \mathbf{S} . We resolve this problem by showing that such a characterization is possible whenever the complement of the region of stability Γ in \mathbb{C}^n which defines \mathbf{S} is a compact *polynomially convex domain*. This generalises the work of Guiver and Bose in [2] on 2-D systems with Γ being the closed unit bidisc \bar{U}^2 . We also answer a question of Guiver [2, Open Prob. 6].

Next the robustness questions in n -D stabilization are studied by specialising the general theory developed in § 2. From there it follows that the stabilizing feedback is

robust in the quotient topology. On the other hand we show that there exist nonstabilizable plants which remain nonstabilizable under arbitrary small perturbations. From this it follows that the set of stabilizable plants, although open, is not dense in the set of all plants.

Finally we wish to emphasise here that an important feature of our theory is that it is in terms of the ideals \mathbf{a} and \mathbf{b} , which are intrinsic to the transfer function p , i.e., they are invariant with respect to the fraction nd^{-1} representing p . This is in contrast to previous work in this subject where the results are stated in terms of some special (usually coprime) fractional representation of p . Our development could therefore be considered the “coordinate-free” approach to stabilizability theory.

2. General theory. Recall from the Introduction the stabilization problem over an integral domain \mathbf{A} ; namely, given a transfer function p in \mathbf{F} the field of fractions of \mathbf{A} , determine those c in \mathcal{F}_p , the fibre over p , which stabilize p . In the following section we develop necessary and sufficient conditions for the solution of this problem.

2.1. General stabilizability conditions. Since p is an element of the field of fractions of an integral domain it corresponds to an equivalence class of pairs in $\mathbf{A} \times \mathbf{A}^*$, i.e., two fractions nd^{-1} and $n_1d_1^{-1}$ represent the same plant p if and only if $nd_1 = n_1d$. Thus in the following, although we sometimes state results in terms of a particular fraction, we need to show, and we do, that they are independent of this representation of p .

Recall from § 1.3.1 the definitions of the ideals \mathbf{a} and \mathbf{b} of \mathbf{A} denoted, respectively, by $(n:d)$ and $(d:n)$. Observe that these ideals are in fact independent of the fraction representing p . For if $p = nd^{-1} = n_1d_1^{-1}$, then

$$\begin{aligned} x \in (n:d) &\Leftrightarrow xd = kn && \text{for some } k \text{ in } \mathbf{A} \\ &\Leftrightarrow xdd_1 = knd_1 && \text{as } d_1 \text{ is nonzero} \\ &\Leftrightarrow xdd_1 = kn_1d \\ &\Leftrightarrow xd_1 = kn_1 && \text{as } d \text{ is nonzero} \\ &\Leftrightarrow x \in (n_1:d_1); \end{aligned}$$

so $\mathbf{a} = (n:d) = (n_1:d_1)$ and similarly for the ideal \mathbf{b} .

Our main result on stabilizability is the following theorem.

THEOREM 2.1.1. *The transfer function p is stabilizable if and only if the ideal quotients of p are coprime, i.e., $\mathbf{a} + \mathbf{b} = \mathbf{A}$.*

Let $p = nd^{-1}$ be any (fixed) representation of p . Then all the stabilizing transfer functions c of p are of the form

$$c = x_3x_1^{-1},$$

where x_1, x_2, x_3 satisfy

$$(2) \quad nx_1 = dx_2,$$

$$(3) \quad nx_3 = d(1 - x_1).$$

Proof. Suppose $p = nd^{-1}$ is stabilized by $c = at^{-1}$ in \mathcal{F}_p . Then by definition there exist x_1, x_2, x_3 in \mathbf{A} such that

$$(4) \quad (1 + pc)^{-1} = x_1,$$

$$(5) \quad p(1 + pc)^{-1} = x_2,$$

$$(6) \quad c(1 + pc)^{-1} = x_3.$$

First, (4) implies that x_1 is nonzero. Next, (5) $\Leftrightarrow px_1 = x_2$ (substituting from (4)) $\Leftrightarrow nx_1 = dx_2$ which is (2). Now (6) $\Leftrightarrow cx_1 = x_3 \Leftrightarrow ax_1 = tx_3$; i.e., $c = at^{-1} = x_3x_1^{-1}$. Finally, (4) $\Leftrightarrow x_1(1+pc) = 1 \Leftrightarrow x_1 + nx_3d^{-1} = 1$ (as $c = x_3x_1^{-1}$ from above) $\Leftrightarrow dx_1 + nx_3 = d$; i.e., $nx_3 = d(1-x_1)$ which is (3). Thus, from (2) $x_1 \in (d:n) = \mathbf{b}$ and from (3) $(1-x_1) \in (n:d) = \mathbf{a}$. Hence, $1 = x_1 + (1-x_1) \in \mathbf{a} + \mathbf{b}$, i.e., \mathbf{a} and \mathbf{b} are coprime.

Conversely, suppose \mathbf{a} and \mathbf{b} are coprime, i.e., suppose there exists x_1 in \mathbf{b} such that $(1-x_1)$ is in \mathbf{a} . We first claim that there exists a *nonzero* x_1 in \mathbf{b} such that $(1-x_1)$ is in \mathbf{a} . For if $\mathbf{a} \neq \mathbf{A}$, then 1 is not in \mathbf{a} . So if $x_1 = 0$ is the only element in \mathbf{b} for which $1-x_1$ is in \mathbf{a} , then 1 is in \mathbf{a} which is absurd. On the other hand suppose $\mathbf{a} = \mathbf{A}$. Now $\mathbf{b} = (d:n)$ is not the zero ideal as d , which is nonzero, belongs to \mathbf{b} . Then for any x_1 in \mathbf{b} , $1-x_1$ is in \mathbf{a} . So for any nonzero x_1 in \mathbf{b} with $1-x_1$ in \mathbf{a} let x_2 and x_3 be elements in \mathbf{A} such that

$$nx_1 = dx_2, \quad nx_3 = d(1-x_1).$$

Let $c = x_3x_1^{-1}$. Then a simple computation shows that

$$(1+pc)^{-1} = x_1, \quad p(1+pc)^{-1} = x_2, \quad c(1+pc)^{-1} = x_3;$$

i.e., p is stabilized by c . \square

In applications it is of interest to determine whether there exist stabilizing controllers for a given plant which are in \mathbf{A} . We call such a plant *strongly stabilizable*. A necessary and sufficient condition for this follows from the above theorem.

COROLLARY 2.1.2. *Let nd^{-1} be any representation of a stabilizable plant p . Then p is strongly stabilizable by a controller c in \mathbf{A} if and only if*

$$(7) \quad (d) \subseteq ((nc+d)).$$

Proof. From the above theorem it follows that the stabilizing controller of p is given by $c = x_3x_1^{-1}$. Hence, c belongs to \mathbf{A} if and only if $x_3 = cx_1 \Leftrightarrow (nc+d)x_1 = d$ (from (3)). \square

DEFINITION. Let \mathbf{T} be a saturated multiplicatively closed subset of \mathbf{A} . A transfer function is said to be *weakly causal* if it belongs to the subring $\mathbf{T}^{-1}\mathbf{A}$ of \mathbf{F} and to be *strictly causal* if it belongs to the Jacobson radical of $\mathbf{T}^{-1}\mathbf{A}$. The subring $\mathbf{T}^{-1}\mathbf{A}$ will be called a *causal structure* defined by \mathbf{T} .

The above definition is motivated by engineering considerations where a plant is weakly causal if its output at time $t=0$ is dependent on the input for time $t \leq 0$, and is strongly causal if it is dependent on the input for time $t < 0$ (see Bose [2]).

PROPOSITION 2.1.3. *All the stabilizing transfer functions of a strictly causal transfer function are weakly causal.*

Proof. Let $p = nd^{-1}$ be a strictly causal transfer function; i.e., let p belong to \mathbf{j} , the Jacobson radical of $\mathbf{T}^{-1}\mathbf{A}$. Hence n belongs to \mathbf{j} . As every stabilizing transfer function $c = x_3x_1^{-1}$ of p satisfies (2) and (3), it follows that $(1-x_1)$ belongs to \mathbf{j} (as d being a unit in $\mathbf{T}^{-1}\mathbf{A}$ does not). Now if x_1 were not to belong to \mathbf{T} , it then would be in a maximal ideal \mathbf{m} of $\mathbf{T}^{-1}\mathbf{A}$ to which also $(1-x_1)$ belongs. Then $x_1 + (1-x_1) = 1$ would belong to \mathbf{m} , which is absurd. \square

2.1.1. Special cases. We now specialize the above general results, which are valid for arbitrary integral domains, to rings of interest in applications.

1. *Bezout domains.* We show here that Theorem 2.1.1 specializes to the well-known result on stabilization (Vidyasagar, Schnider, and Francis [10]).

COROLLARY 2.1.4. *If \mathbf{A} is a Bezout domain then every transfer function is stabilizable.*

Proof. Recall that in a Bezout domain every finitely generated ideal is principal. So for $p = nd^{-1}$, let $(n, d) = (g)$ for some g in \mathbf{A} where (g) is the smallest ideal that contains both (n) and (d) . Hence $n = n'g$ and $d = d'g$ for some n', d' in \mathbf{A} .

We claim that $(n', d') = \mathbf{A}$. For if not then (n', d') is a proper ideal in \mathbf{A} and hence equals (g') for some g' not a unit. Then (gg') is a principal ideal strictly contained in (g) and which contains (n) and (d) , which is a contradiction.

Clearly $(n') \subseteq (n': d') = (n: d) = \mathbf{a}$; similarly $(d') \subseteq \mathbf{b}$. Hence, $\mathbf{a} + \mathbf{b} = \mathbf{A}$. \square

Remark. In general, even if \mathbf{A} is not a Bezout domain, the above argument shows that if p can be represented by nd^{-1} where (n) and (d) are coprime then p is stabilizable.

2. *Unique factorization domains.* These rings are of interest in distributed systems and n -D problems. For instance the rings $\mathbb{K}[X_1 \cdots X_n]$ and $\mathbb{K}[[X_1 \cdots X_n]]$ are UFDs but not Bezout domains. In this case for $p = nd^{-1}$ there exist relatively prime n' and d' such that $p = n'd'^{-1}$. Then clearly $(n: d) = (n')$ and $(d: n) = (d')$. Hence we have the following corollary.

COROLLARY 2.1.5. *Let \mathbf{A} be a UFD. Then p is stabilizable if and only if*

$$(n') + (d') = \mathbf{A}.$$

Remark. For a general ring \mathbf{A} , it follows from (3) that every stabilizing controller of $p = nd^{-1}$ is of the form $c = x_3x_1^{-1}$ where x_3 and x_1 are elements of the ring \mathbf{A} that satisfy

$$nx_3 + dx_1 = d.$$

In the special case of a UFD if n and d are relatively prime, it follows from (2) and (3) that

$$x_1 = td \quad \text{and} \quad x_3 = ad.$$

Hence the above equation becomes

$$nad + dtd = d,$$

which, as d is nonzero, is equivalent to

$$na + dt = 1.$$

Thus every stabilizing controller c of $p = nd^{-1}$ (n and d relatively prime) is of the form $c = at^{-1}$ where a and t are elements of \mathbf{A} that solve $na + dt = 1$ or equivalently $na + dt = u$, for u a unit in \mathbf{A} .

With reference to the strong stabilizability problem described above we have the following corollary.

COROLLARY 2.1.6. *Let \mathbf{A} be a UFD. Then a stabilizable transfer function $p = n'd'^{-1}$, where n' and d' are relatively prime, is strongly stabilizable by a c in \mathbf{A} if and only if*

$$((n'c + d')) = \mathbf{A}.$$

Proof. From (2) above we get $x_1 = yd'$ for some y in \mathbf{A} (since n' and d' are relatively prime). Hence, from (3) and the fact that $x_3 = cx_1$ (which is equivalent to the strong stabilizability of p) we have $(n'c + d')yd' = d' \Leftrightarrow (n'c + d')$ is a unit. \square

3. *Noetherian rings.* We now consider the case when \mathbf{A} is a Noetherian ring since this ring is of importance to many applications. As there exist many Noetherian rings that are not UFDs this case merits separate treatment.

Consider the primary decompositions of (n) and (d) ,

$$(n) = \bigcap_{i=1}^{m_1} \mathbf{q}_i \quad \text{and} \quad (d) = \bigcap_{j=1}^{m_2} \mathbf{q}'_j,$$

where the indices i and j are ordered such that

$$d \notin \mathbf{q}_i \quad \text{for } i = 1 \cdots k_1 \leq m_1$$

and

$$n \notin \mathbf{q}'_j \quad \text{for } j = 1 \cdots k_2 \leq m_2.$$

Let $\text{rad } \mathbf{q}_i = \mathbf{p}_i$ and $\text{rad } \mathbf{q}'_j = \mathbf{p}'_j$. (Here rad denotes the radical.) In this notation we have the following corollary.

COROLLARY 2.1.7. *Let \mathbf{A} be a Noetherian ring. Then $p = nd^{-1}$ is stabilizable if and only if*

$$\mathbf{p}_i + \mathbf{p}'_j = \mathbf{A} \quad \forall i = 1 \cdots k_1, \quad j = 1 \cdots k_2.$$

Proof. Since \mathbf{a} and \mathbf{b} coprime is equivalent to $\text{rad } \mathbf{a}$ and $\text{rad } \mathbf{b}$ coprime, p is stabilizable if and only if

$$(8) \quad \text{rad } \mathbf{a} + \text{rad } \mathbf{b} = \mathbf{A}.$$

Since for a primary ideal \mathbf{q} and an element x in \mathbf{A}

$$\begin{aligned} \text{rad } (\mathbf{q} : x) &= \text{rad } \mathbf{q} \quad \text{if } x \notin \mathbf{q} \\ &= \mathbf{A} \quad \text{if } x \in \mathbf{q} \end{aligned}$$

we have

$$(9) \quad \text{rad } \mathbf{a} = \text{rad } \bigcap_{i=1}^{m_1} (\mathbf{q}_i : d) = \bigcap_{i=1}^{k_1} \mathbf{p}_i$$

and similarly

$$\text{rad } \mathbf{b} = \bigcap_{j=1}^{k_2} \mathbf{p}'_j.$$

Hence (8) implies that $\mathbf{p}_i + \mathbf{p}'_j = \mathbf{A}$ for all $i = 1 \cdots k_1$ and $j = 1 \cdots k_2$.

Conversely suppose p is not stabilizable. Then the ideal

$$\bigcap_{i=1}^{k_1} \mathbf{p}_i + \bigcap_{j=1}^{k_2} \mathbf{p}'_j \neq \mathbf{A},$$

and hence is contained in some prime ideal \mathbf{p} of \mathbf{A} . Thus, $\bigcap_{i=1}^{k_1} \mathbf{p}_i \subseteq \mathbf{p}$ and $\bigcap_{j=1}^{k_2} \mathbf{p}'_j \subseteq \mathbf{p} \Rightarrow \mathbf{p}_i \subseteq \mathbf{p}$ for some $i \leq k_1$ and $\mathbf{p}'_j \subseteq \mathbf{p}$ for some $j \leq k_2 \Leftrightarrow \mathbf{p}_i + \mathbf{p}'_j \subseteq \mathbf{p}$. \square

Remark. Note that for a different fraction $n'd'^{-1}$ representing the transfer function p the associated prime ideals of (n') and (d') will in general be different from those of (n) and (d) . However, in light of the comment preceding Theorem 2.1.1 the intersections in (9) will be independent of these representations and so will therefore be the condition of the above corollary.

4. Ring of fractions. Now consider the case when \mathbf{A} itself arises as a ring of fractions of some other ring \mathbf{B} , i.e., $\mathbf{A} = \mathbf{S}^{-1}\mathbf{B}$, $\mathbf{S} \subseteq \mathbf{B}$ is a multiplicatively closed subset (not containing 0). This case is of importance to applications and one such application, namely the multidimensional stabilization problem, is considered later in this paper. There \mathbf{B} is the polynomial ring $\mathbb{C}[X_1 \cdots X_n]$ and \mathbf{S} is a set of polynomials whose varieties do not intersect some fixed region Γ in \mathbb{C}^n . We now interpret our basic stabilization theorem in terms of ideals of \mathbf{B} .

Let \mathbf{I}_a and \mathbf{I}_b be ideals of \mathbf{B} such that

$$\mathbf{a} = \mathbf{S}^{-1}\mathbf{I}_a \quad \text{and} \quad \mathbf{b} = \mathbf{S}^{-1}\mathbf{I}_b.$$

Theorem 2.1.1 now translates to the following corollary.

COROLLARY 2.1.8. *The transfer function p is stabilizable if and only if*

$$(10) \quad (\mathbf{I}_a + \mathbf{I}_b) \cap \mathbf{S} \neq \emptyset.$$

Proof. As $\mathbf{a} + \mathbf{b} = \mathbf{S}^{-1}(\mathbf{I}_a + \mathbf{I}_b)$ it follows that

$$\mathbf{a} + \mathbf{b} = \mathbf{A} \Leftrightarrow (\mathbf{I}_a + \mathbf{I}_b) \cap \mathbf{S} \neq \emptyset. \quad \square$$

Note that \mathbf{I}_a and \mathbf{I}_b need not be unique. However, we determine one such pair in terms of a suitable fraction representing p .

Let $p = nd^{-1}$, n, d in $\mathbf{S}^{-1}\mathbf{B}$, i.e., $n = f'h'^{-1}$ and $d = g'h''^{-1}$ for h', h'' in \mathbf{S} . Then $p = nd^{-1} = (f'h''/1)(g'h'/1)^{-1}$. This is well-defined for as h'' is in \mathbf{S} , $h''/1$ is nonzero. Hence $g'/1 = (g'h''^{-1})(h''/1)$ is nonzero, which implies that $g'h'/1$ is nonzero.

Let $f = f'h''$, $g = g'h'$ so that

$$(11) \quad p = (f/1)(g/1)^{-1}.$$

In this notation we have the following proposition.

PROPOSITION 2.1.9. $\mathbf{I}_a = (f: g)$ and $\mathbf{I}_b = (g: f)$ extend to \mathbf{a} and \mathbf{b} , respectively.

Proof. Let $h \in (f: g)$. Then $(g/1)(h/1)$ belongs to the ideal $(f/1)$, which implies that $h/1$ is in \mathbf{a} . (This follows from the representation of p in (11).) Thus $\mathbf{S}^{-1}\mathbf{I}_a \subseteq \mathbf{a}$.

Conversely let a/s be in \mathbf{a} . Then

$$(a/s)(g/1) = (a'/s')(f/1)$$

for some a'/s' in \mathbf{A} . This implies that

$$as' \in (f: g) \Rightarrow as'/(s's) = a/s \in \mathbf{S}^{-1}\mathbf{I}_a,$$

i.e., $\mathbf{a} \subseteq \mathbf{S}^{-1}\mathbf{I}_a$.

The proof for $\mathbf{b} = \mathbf{S}^{-1}\mathbf{I}_b$ is identical. \square

Remark. In the case where \mathbf{B} is a UFD there exist f and g in \mathbf{B} that are relatively prime such that $(f/1)(g/1)^{-1} = p$. Then $\mathbf{I}_a = (f)$ and $\mathbf{I}_b = (g)$. Hence condition (10) is equivalent to

$$(12) \quad p \text{ is stabilizable iff } (f, g) \cap \mathbf{S} \neq \emptyset.$$

Remark. Causal structures in this case will be defined via saturated multiplicatively closed subsets \mathbf{T} of $\mathbf{A} = \mathbf{S}^{-1}\mathbf{B}$ which contain the image of \mathbf{S} under the natural injection $i: \mathbf{B} \rightarrow \mathbf{S}^{-1}\mathbf{B}$. Then $\mathbf{T}^{-1}\mathbf{A}$ will be naturally identified with $\mathbf{T}^{-1}\mathbf{B}$.

2.2. Geometric interpretations. In this section we interpret the above algebraic criteria of stabilizability in concrete geometric terms. Recall from § 1.3.2 that $\text{spec } \mathbf{A}$ denotes the prime spectrum of the ring \mathbf{A} and for an ideal \mathbf{c} in \mathbf{A} , $V(\mathbf{c})$ denotes the subset of $\text{spec } \mathbf{A}$ consisting of all prime ideals that contain \mathbf{c} . Clearly the geometric equivalent of Theorem 2.1.1 is the following corollary.

COROLLARY 2.2.10. *The transfer function p is stabilizable if and only if*

$$V(\mathbf{a}) \cap V(\mathbf{b}) = \emptyset.$$

The equivalents of the above geometric interpretation for the special cases of Bezout domains, UFDs, and Noetherian rings are straightforward. We consider therefore only the case when \mathbf{A} is a ring of fractions $\mathbf{S}^{-1}\mathbf{B}$ and determine the geometric equivalent of Corollary 2.1.8.

Corresponding to the multiplicatively closed subset \mathbf{S} of \mathbf{B} , define

$$\Omega = \{\mathbf{p} \in \text{spec } \mathbf{B} \mid \mathbf{p} \cap \mathbf{S} = \emptyset\}.$$

In this case we have the following corollary.

COROLLARY 2.2.11. *The transfer function p is stabilizable if and only if*

$$V(\mathbf{I}_a) \cap V(\mathbf{I}_b) \cap \Omega = \emptyset.$$

Proof. Assume p stabilizable, which implies by Corollary 2.1.8 that

$$(\mathbf{I}_a + \mathbf{I}_b) \cap \mathbf{S} \neq \emptyset \Rightarrow \mathbf{p} \cap \mathbf{S} \neq \emptyset \quad \text{for all } \mathbf{p} \in V(\mathbf{I}_a + \mathbf{I}_b) \Rightarrow V(\mathbf{I}_a) \cap V(\mathbf{I}_b) \cap \Omega = \emptyset.$$

Conversely suppose p is not stabilizable. Then again by Corollary 2.1.8 $(\mathbf{I}_a + \mathbf{I}_b) \cap \mathbf{S} = \emptyset$, which implies that $\mathbf{S}^{-1}(\mathbf{I}_a + \mathbf{I}_b)$ is a proper ideal in $\mathbf{S}^{-1}\mathbf{B}$. Let \mathbf{m} be a maximal ideal in $\mathbf{S}^{-1}\mathbf{B}$ containing $\mathbf{S}^{-1}(\mathbf{I}_a + \mathbf{I}_b)$. Let \mathbf{p} and \mathbf{i} be the contractions in \mathbf{B} of ideals \mathbf{m} and $\mathbf{S}^{-1}(\mathbf{I}_a + \mathbf{I}_b)$, respectively. Then $(\mathbf{I}_a + \mathbf{I}_b) \subseteq \mathbf{i} \subseteq \mathbf{p}$. Since $\mathbf{p} \cap \mathbf{S} = \emptyset$ and \mathbf{p} belongs to $V(\mathbf{I}_a + \mathbf{I}_b)$, the result follows. \square

Observe that Corollary 2.2.10 requires the determination of the (Zariski) closed subsets $V(\mathbf{a})$ and $V(\mathbf{b})$ in terms of the given data $V(n)$ and $V(d)$. Note that whenever the plant can be represented by a relatively prime fraction nd^{-1} (for instance whenever \mathbf{A} is either a Bezout domain or a UFD) $V(\mathbf{a}) = V(n)$ and $V(\mathbf{b}) = V(d)$. Also in the Noetherian case it is clear from (9) in Corollary 2.1.7 that

$$V(\mathbf{a}) = \bigcup_{i=1}^{k_1} V(\mathbf{p}_i) \quad \text{and} \quad V(\mathbf{b}) = \bigcup_{j=1}^{k_2} V(\mathbf{p}'_j).$$

Thus again $V(\mathbf{a})$ and $V(\mathbf{b})$ are expressed in terms of n and d (i.e., in terms of the primary decompositions of (n) and (d)). Note that similar expressions can be derived for $V(\mathbf{I}_a)$ and $V(\mathbf{I}_b)$ in Corollary 2.2.11 when \mathbf{A} is a ring of fractions of a Noetherian ring \mathbf{B} . However, while in general there does not seem to be such a simple way of expressing $V(\mathbf{a})$ and $V(\mathbf{b})$, we show below that for certain transfer functions which we call *simple* a geometric characterization of $V(\mathbf{a})$ and $V(\mathbf{b})$ is indeed possible.

DEFINITION. A transfer function p is said to be *simple* if it has a representation nd^{-1} with $\text{rad}(n) = (n)$ and $\text{rad}(d) = (d)$. We call such a representation simple.

Remark. This definition is motivated by the fact that in many practical situations transfer functions of plants have simple (i.e., nonrepeated) nonminimum phase poles and zeros. Thus in view of our remarks in § 1.3.2 concerning the prime spectrum of the ring the above definition is a natural generalization of transfer functions with simple poles and zeros. Note that in the special case of a simple n -D plant fg^{-1} described above in subsection 4 of § 2.1, those irreducible factors of the polynomials f and g that do not belong to \mathbf{S} are nonrepeated. In fact it can be shown that in the *quotient topology* defined in the next section simple transfer functions are open and dense in the set of all transfer functions, i.e., most transfer functions are in fact simple.

For such transfer functions we have the following proposition.

PROPOSITION 2.2.12. *Let p have a simple representation nd^{-1} . Define Γ_a and Γ_b to be the families of closed subsets of $\text{spec } \mathbf{A}$ given by*

$$\Gamma_a = \{\psi \mid V(n) \subseteq \psi \cup V(d)\},$$

$$\Gamma_b = \{\theta \mid V(d) \subseteq \theta \cup V(n)\}.$$

Then $V(\mathbf{a})$ and $V(\mathbf{b})$ are the unique minimal elements of Γ_a and Γ_b , respectively.

Proof. First we show that $V(\mathbf{a})$ belongs to Γ_a . So let x be in \mathbf{a} , which implies that $(x)(d) \subseteq (n)$. Hence

$$V(n) \subseteq V(x) \cup V(d) \quad \text{for all } x \in \mathbf{a} \Rightarrow V(n) \subseteq \left(\bigcap_{x \in \mathbf{a}} V(x) \right) \cup V(d) = V(\mathbf{a}) \cup V(d).$$

Thus $V(\mathbf{a})$ belongs to Γ_a .

We claim that if ψ belongs to Γ_a then $V(\mathbf{a}) \subseteq \psi$. For as ψ is closed there exists an ideal \mathbf{i} in \mathbf{A} such that $\psi = V(\mathbf{i})$. Then

$$\begin{aligned} V(n) \subseteq V(\mathbf{i}) \cup V(d) = V(\mathbf{i}(d)) &\Rightarrow \mathbf{i}(d) \subseteq \text{rad}(n) = (n) \quad (\text{as } p \text{ is simple}) \\ &\Rightarrow \mathbf{i} \subseteq (n : d) = \mathbf{a} \Rightarrow V(\mathbf{a}) \subseteq V(\mathbf{i}). \end{aligned}$$

A similar reasoning shows that $V(\mathbf{b})$ is minimal in Γ_b . \square

Thus it clearly follows that the geometric stabilizability condition of Corollary 2.1.10 can now be expressed for simple transfer functions as

$$p \text{ is stabilizable iff } \psi_{\min} \cap \theta_{\min} = \emptyset,$$

where ψ_{\min} and θ_{\min} are minimal elements of the families Γ_a and Γ_b , respectively.

2.3. Robustness of stabilizability. In this section we investigate the robustness of feedback stability over general rings. For this purpose we need a notion of when one transfer function is close to another, i.e., we need to topologize the set of transfer functions. In the special case of a Bezout domain the graph topology defined by Vidyasagar [8] serves this purpose. However, since this topology is defined in terms of coprime fractional representations of the transfer function it cannot be carried over to our more general setting. Our first purpose therefore is to topologize the set of transfer functions in such a way so as to be independent of any choice of fractional representations. This is in keeping with the spirit of this paper as the above developments have been in terms of the ideals \mathbf{a} and \mathbf{b} , which are indeed independent of the representation.

Motivated by the above considerations we topologize the set of transfer functions in this general setting in such a way so as to specialise to the graph topology whenever coprime fractions exist. Our development here is also guided by genericity questions in the multidimensional stabilization problem, which is treated in detail in the next section.

Now let the ring \mathbf{A} be a topological ring (\mathbf{A}, τ_R) . The field of fractions \mathbf{F} arises as a quotient of $\mathbf{A} \times \mathbf{A}^*$ by the equivalence relation \sim where

$$(a, t) \sim (a', t') \quad \text{iff} \quad at' = a't.$$

Consider $\mathbf{A} \times \mathbf{A}$ to be a topological space with some topology τ (not necessarily the product topology). We consider \mathbf{F} to be a topological space with the *quotient topology* τ_q induced by the subspace topology, also denoted by τ , on $\mathbf{A} \times \mathbf{A}^* \subset \mathbf{A} \times \mathbf{A}$, i.e., if

$$(13) \quad \pi: \mathbf{A} \times \mathbf{A}^* \rightarrow \mathbf{F}$$

is the natural projection, then $U \subseteq \mathbf{F}$ is open in τ_q if and only if $\pi^{-1}(U)$ belongs to τ . Our choice of the topology τ on $\mathbf{A} \times \mathbf{A}$ is motivated by genericity questions and is required to satisfy the following conditions. These conditions as well as others that we introduce in this section will be shown to be satisfied in the special case of n -D systems in the next section.

Condition C1. Every element of \mathbf{A} is closed in τ_R .

Condition C2. Let $ay = bx$. Then given a neighborhood $N(x, y)$ of (x, y) (i.e., in τ) there exists a neighborhood $N(a, b)$ of (a, b) such that for all (a', b') in $N(a, b)$; there exists (x', y') in $N(x, y)$ with $a'y' = b'x'$.

Condition C3. The product topology τ_P on $\mathbf{A} \times \mathbf{A}$ is weaker than the topology τ on $\mathbf{A} \times \mathbf{A}$, i.e., $\tau_P \subseteq \tau$.

Remark. Clearly under assumptions C1 and C3, $\mathbf{A} \times \mathbf{A}^*$ is open in $\mathbf{A} \times \mathbf{A}$ with respect to the topology τ . Then Condition C2 is satisfied by the subspace $\mathbf{A} \times \mathbf{A}^*$, i.e., if (a, b) and (x, y) belong to $\mathbf{A} \times \mathbf{A}^*$ then the neighborhoods $N(a, b)$ and $N(x, y)$ in C2 can be chosen to be open subsets of $\mathbf{A} \times \mathbf{A}^*$. We then say that $\mathbf{A} \times \mathbf{A}^*$ satisfies C2.

Remark. It is important to note that τ , in general, is not the product topology τ_P . The open sets of τ_P are usually much too large to allow robustness of stabilizability. This will become clear in the section on n -D systems below.

PROPOSITION 2.3.13. *The mapping π in (13) is open if and only if $\mathbf{A} \times \mathbf{A}^*$ satisfies C2.*

Proof. Suppose π is open. Then given a neighborhood $N(x, y)$ of (x, y) , $\pi N(x, y)$ is open, which implies that $\pi^{-1}\pi N(x, y)$ is open (as \mathbf{F} has the quotient topology).

Let $(a, b) \sim (x, y)$, i.e., $(a, b) \in \pi^{-1}\pi N(x, y)$. As $\pi^{-1}\pi N(x, y)$ is open, there exists a neighborhood $N(a, b)$ of (a, b) contained in $\pi^{-1}\pi N(x, y)$. Hence for all (a', b') in $N(a, b)$ there exists (x', y') in $N(x, y)$ such that $(a', b') \sim (x', y')$, i.e., C2 is satisfied.

Conversely suppose that $\mathbf{A} \times \mathbf{A}^*$ satisfies C2. Let $N \subseteq \mathbf{A} \times \mathbf{A}^*$ be open. We need to show that $\pi^{-1}\pi N$ is open. So let (a, b) be in $\pi^{-1}\pi N$. Clearly there exists (x, y) in N such that $(a, b) \sim (x, y)$. By C2 there exists a neighborhood $N(a, b)$ of (a, b) such that for all (a', b') in $N(a, b)$, there exists (x', y') in N with $(a', b') \sim (x', y')$. Hence $N(a, b) \subseteq \pi^{-1}\pi N$. \square

Remark. Clearly by the above proposition a basis of neighborhoods containing a transfer function p in \mathbf{F} can be obtained as follows.

Let (a, b) in $\mathbf{A} \times \mathbf{A}^*$ belong to $\pi^{-1}p$. Consider the collection πU , where U varies over all basic neighborhoods of (a, b) in $\mathbf{A} \times \mathbf{A}^*$ and (a, b) varying over $\pi^{-1}p$. Then this collection is a basis of neighborhoods about p . Moreover, it follows from C2 that this basis can actually be obtained by fixing any (a, b) in $\pi^{-1}p$.

From this it follows that whenever every p has coprime fractions, the quotient topology on \mathbf{F} is just the graph topology.

We now investigate the following questions:

1. Given a stabilizable transfer function p is there a neighborhood in the quotient topology defined above consisting of stabilizable transfer function?

2. Given such a neighborhood does there exist a controller that stabilizes every transfer function in it?

Consider the topological ring (\mathbf{A}, τ_R) with a topology τ on $\mathbf{A} \times \mathbf{A}$ satisfying the above conditions C1, C2, and C3. Assume further that the following condition holds.

Condition C4. The set of units in \mathbf{A} is open in the topology τ_R .

Under Conditions C1 to C4 we have the following proposition.

PROPOSITION 2.3.14. *Let p in \mathbf{F} be stabilizable. Then there exists a neighborhood of p in the quotient topology such that every transfer function in this neighborhood is stabilizable.*

Proof. Let $p = nd^{-1}$. Then by Theorem 2.1.1 p is stabilizable if and only if there exist x_1, x_2, x_3 in \mathbf{A} such that the following equations hold:

$$nx_1 = dx_2, \quad nx_3 = d(1 - x_1).$$

Observe that $x_1 + (1 - x_1) = 1$ is contained in the open set of units. Hence as \mathbf{A} is a topological ring there exist neighborhoods $N_P(x_2, x_1)$ and $N_P(1 - x_1, x_3)$ of (x_2, x_1)

and $(1 - x_1, x_3)$ in the product topology τ_P such that for all (x'_2, x'_1) in $N_P(x_2, x_1)$ and (x'_4, x'_3) in $N_P(1 - x_1, x_3)$

$$(14) \quad x'_1 + x'_4 = u$$

where u is a unit.

Now choose neighborhoods $N(x_2, x_1)$ and $N(1 - x_1, x_3)$ in the topology τ on $\mathbf{A} \times \mathbf{A}$ such that

$$N(x_2, x_1) \subseteq N_P(x_2, x_1)$$

and

$$N(1 - x_1, x_3) \subseteq N_P(1 - x_1, x_3)$$

which is possible by Condition C3. Given these neighborhoods, by C2 there exists a neighborhood $N(n, d)$ in $\mathbf{A} \times \mathbf{A}^*$ such that for all (n', d') in $N(n, d)$, there exist (x'_2, x'_1) in $N(x_2, x_1)$ and (x'_4, x'_3) in $N(1 - x_1, x_3)$ with

$$n'x'_1 = d'x'_2,$$

$$n'x'_3 = d'x'_4.$$

But by (14) $u^{-1}x'_1 + u^{-1}x'_4 = 1$. Hence multiplying both the above equations by u^{-1} we obtain

$$n'(x'_1u^{-1}) = d'(x'_2u^{-1}),$$

$$n'(x'_3u^{-1}) = d'(1 - x'_1u^{-1}).$$

Hence, $n'd'^{-1}$ is also stabilizable for all (n', d') in $N(n, d)$. \square

While this proposition answers question 1 above, the controller $x_3x_1^{-1}$ of p may not be the same as the stabilizing controller $x'_3x'_1{}^{-1}$ of $n'd'^{-1}$. We answer question 2 in the affirmative in the special case when \mathbf{A} is a UFD satisfying Conditions C1–C4.

THEOREM 2.3.15. *Let \mathbf{A} be a UFD satisfying C1–C4. Suppose c stabilizes a transfer function p in \mathbf{F} . Then there exists a neighborhood of p in the quotient topology such that c stabilizes every plant in it.*

Proof. Let $p = nd^{-1}$ where now n and d are relatively prime. Then from (2) and (3) of Theorem 2.1.1 with these n and d , $c = x_3x_1^{-1}$. Furthermore,

$$x_1 = rd \quad \text{and} \quad 1 - x_1 = sn$$

for some r, s in \mathbf{A} .

By Conditions C3 and C4 there exists a neighborhood $N(n, d)$ of (n, d) in the topology τ on $\mathbf{A} \times \mathbf{A}$ such that for all (n', d') in $N(n, d)$

$$rd' + sn' = u,$$

where u is a unit in \mathbf{A} .

Define

$$x'_1 = rd'u^{-1}, \quad x'_2 = rn'u^{-1}, \quad x'_3 = sd'u^{-1}, \quad \text{and} \quad x'_4 = sn'u^{-1}.$$

Then

$$n'x'_1 = d'x'_2, \quad n'x'_3 = d'(1 - x'_1).$$

Thus $n'd'^{-1}$ is stabilized by $x'_3x'_1{}^{-1} = x_3x_1^{-1} = c$. \square

While the above theorem answers question 2 in the special case of a UFD, the corresponding answer for the general case is not known.

By the above the set of stabilizable transfer functions is open in the quotient topology. Is it also dense in the set of all transfer functions? Although the answer to this is in the affirmative in the case of LTI systems (as is well known), this is not so in the more general setting of this paper. In fact we show in the special case of n -D systems in the next section that there exist nonstabilizable transfer functions with neighborhoods consisting of transfer functions that are also not stabilizable.

For related results regarding the graph topology see Zhu [11].

3. Stabilization of multidimensional systems. In this section we use the above theory to analyse stabilizability of multidimensional systems.

DEFINITION OF THE PROBLEM. With respect to the general formulation of the stabilization problem in the introduction, the problem here is defined as follows:

Let $\mathbf{A} = \mathbf{S}^{-1}\mathbf{B}$ where $\mathbf{B} = \mathbb{K}[X_1 \cdots X_n]$, the polynomial ring in n indeterminates with coefficients in a subfield \mathbb{K} of \mathbb{C} .

Let \mathbf{S} be the multiplicatively closed subset of \mathbf{B} consisting of all polynomials whose (affine) varieties in \mathbb{C}^n do not intersect some fixed compact region $\Gamma \subset \mathbb{C}^n$.

We define a causal structure via closed subsets of Γ as follows:

Let Γ' be some fixed closed subset of Γ .

Let \mathbf{T} be the saturated multiplicatively closed subset of \mathbf{A} consisting of rational functions fg^{-1} such that the variety of f does not intersect Γ' . (Note that the natural injection $i: \mathbf{B} \rightarrow \mathbf{A}$ maps \mathbf{S} into \mathbf{T} ; viz., the remark at the very end of § 2.1.)

Thus this is a special case of the general problem considered in part 4 of § 2.1, namely when \mathbf{A} is a ring of fractions. Also observe that as \mathbf{B} is a UFD so is $\mathbf{S}^{-1}\mathbf{B}$.

The main purpose of this section is to interpret the results developed above in this concrete setting of the polynomial ring $\mathbb{K}[X_1 \cdots X_n]$ in terms of affine varieties in \mathbb{C}^n .

The motivation for considering this special case stems from the 2-D stabilization problem treated by Guiver and Bose in [2, Chap. 3]. In our formulation their problem is as follows:

$\mathbf{B} = \mathbb{C}[X_1, X_2]$, $\Gamma = \bar{U}^2$ the closed unit polydisc in \mathbb{C}^2 . The causal structure is defined by $\Gamma' = (0, 0)$.

They prove that a plant $p = f/g$, f and g in $\mathbb{C}[X_1, X_2]$ is stabilizable if and only if f and g do not have a common zero in \bar{U}^2 . Their proof makes critical use of the fact that f and g , when relatively prime, have a finite number of common zeros (which is not the case for n -D systems, $n > 2$), and that the region Γ defining \mathbf{S} is the unit polydisc.

We on the other hand treat n -D systems for general n and where the region Γ is any arbitrary compact region in \mathbb{C}^n . The motivation for generalizing from 2-D to n -D is clear. Replacing the polydisc \bar{U}^n by an arbitrary compact region Γ follows from applications and is in fact posed as an *open problem* by Guiver in [2, Open Prob. 6].

In the following section we arrive at geometric criteria based on the theory developed in the previous sections and which also allows us to settle this open problem.

3.1. Stabilizability conditions. Consider the n -D transfer function p in \mathbf{F} . Since \mathbf{B} here is a UFD, by the remark following Proposition 2.1.9 p admits a representation $(f/1)(g/1)^{-1}$, where f and g are relatively prime. Then $\mathbf{I}_a = (f)$ and $\mathbf{I}_b = (g)$. Recall also from § 2.2 that Ω is the set of all prime ideals of \mathbf{B} that do not intersect \mathbf{S} . Hence, by Corollary 2.2.11, we have

$$(15) \quad p \text{ is stabilizable iff } V(f, g) \cap \Omega = \emptyset.$$

Since Ω is a set of prime ideals of $\mathbb{K}[X_1 \cdots X_n]$, and since by the Hilbert Nullstellensatz these prime ideals correspond to irreducible varieties in \mathbb{C}^n (as \mathbb{C} is the algebraic closure of \mathbb{K}), we can think of Ω as a collection of irreducible varieties. (As our entire discussion is with respect to the fixed subfield \mathbb{K} of \mathbb{C} , by variety we always mean a \mathbb{K} -variety, i.e., the zero set in \mathbb{C}^n of an ideal in $\mathbb{K}[X_1 \cdots X_n]$.) Now the affine variety of the ideal (f, g) , also denoted $V(f, g)$, is a (finite) union of irreducible varieties as well (corresponding to the minimal primes belonging to the ideal (f, g)). Hence (15) above can be interpreted to mean that none of the irreducible varieties in $V(f, g)$ belongs to Ω considered now as a collection of irreducible varieties.

This interpretation of the stabilizability condition in terms of affine varieties forces on us the need to characterize the irreducible varieties in the collection Ω . In general, since Ω is just a set of prime ideals which do not intersect some given multiplicatively closed subset S , such a characterization may not be possible. However, note that in the above stabilizability problem S arises in a special way, namely as a set of polynomials whose varieties do not intersect some given region $\Gamma \subset \mathbb{C}^n$. For such S we show that a geometric characterization of Ω is indeed possible, and the nature of this characterization is suggested by the following condition.

LEMMA 3.1.16. *Let \mathfrak{p} be a prime ideal in $\mathbb{K}[X_1 \cdots X_n]$. Then*

$$V(\mathfrak{p}) \cap \Gamma \neq \emptyset \Rightarrow \mathfrak{p} \in \Omega.$$

Proof. The proof is obvious. (Here $V(\mathfrak{p})$ is the variety of \mathfrak{p} in \mathbb{C}^n .) \square

The question therefore is: under what conditions on Γ does every \mathfrak{p} in Ω satisfy $V(\mathfrak{p}) \cap \Gamma \neq \emptyset$? Note that for such Γ the stabilizability condition would reduce to the following:

The transfer function $p = fg^{-1}$ (f, g relatively prime) is stabilizable if and only if the varieties $V(f)$ and $V(g)$ do not intersect in Γ .

Remark. Note that as f and g are relatively prime no irreducible component of $V(f)$ coincides with an irreducible component of $V(g)$. Hence, the dimension of every irreducible component of $V(f) \cap V(g)$ is strictly less than $n - 1$. On the other hand as $\text{codimension } V(f) + \text{codimension } V(g) \geq \text{codimension } (V(f) \cap V(g))$ it follows that the dimension of every irreducible component of $V(f) \cap V(g)$ is greater than or equal to $n - 2$. Hence the dimension of every irreducible component of $V(f) \cap V(g)$ equals $n - 2$, i.e., $V(f) \cap V(g)$ is a variety of pure codimension 2.

Thus actually it suffices, as far as the stabilizability condition is concerned, to obtain properties of Γ under which the reverse implication in Lemma 3.1.16 holds for prime ideals of height 2 in Ω .

To repeat, the question now is that if $V(\mathfrak{p}) \cap \Gamma = \emptyset$, then is it true that $\mathfrak{p} \notin \Omega$, i.e., is $\mathfrak{p} \cap S \neq \emptyset$. If this were true then there is a polynomial f in $\mathfrak{p} \cap S$. But

$$f \in S \Rightarrow V(f) \cap \Gamma = \emptyset \quad \text{and} \quad f \in \mathfrak{p} \Rightarrow V(\mathfrak{p}) \subseteq V(f)$$

in which case there is a codimension 1 variety, namely $V(f)$, that contains $V(\mathfrak{p})$ and that does not intersect Γ . This motivates the following.

DEFINITION. A region $\Gamma \subset \mathbb{C}^n$ is said to be *codimension k -convex* if given a codimension k irreducible variety V with $V \cap \Gamma = \emptyset$, there exists a codimension 1 variety V' such that

$$(16) \quad V \subset V' \quad \text{and} \quad V' \cap \Gamma = \emptyset.$$

Remark. If the region Γ is codimension k -convex then clearly for any variety V of pure codimension k with $V \cap \Gamma = \emptyset$ there exists a codimension 1 variety V' such that (16) holds.

We now show that the codimension k -convexity of the region Γ allows the reverse implication in Lemma 3.1.16 for primes of height $n - k$.

PROPOSITION 3.1.17. *Let \mathbf{i} be the ideal in $\mathbb{K}[X_1 \cdots X_n]$ of a pure codimension k variety V in \mathbb{C}^n . Then*

$$(17) \quad V \cap \Gamma = \emptyset \Leftrightarrow \mathbf{i} \cap \mathbf{S} \neq \emptyset \quad \text{iff } \Gamma \text{ is codimension } k\text{-convex.}$$

Proof. That (17) implies codimension k -convexity of Γ is clear. So suppose now that Γ is codimension k -convex. Then given the variety V of the ideal \mathbf{i} with $V \cap \Gamma = \emptyset$, by the remark following the definition above, there exists a codimension 1 variety $V' = V(h)$, where h is a polynomial in $\mathbb{K}[X_1 \cdots X_n]$, such that $V \subset V'$ and $V' \cap \Gamma = \emptyset$. By the Nullstellensatz, h^t belongs to \mathbf{i} for some $t \geq 1$. Since $V' = V(h')$, $V' \cap \Gamma = \emptyset$ implies that h' belongs to $\mathbf{S} \cap \mathbf{i}$. \square

Specialising the above result for $k = 2$ we have the geometric equivalent of the stabilizability result.

THEOREM 3.1.18. *Let Γ be codimension 2-convex. Then the transfer function $p = fg^{-1}$ (where f and g are relatively prime) is stabilizable if and only if*

$$V(f) \cap V(g) \cap \Gamma = \emptyset.$$

The above property of codimension k -convexity of a region Γ is a geometric one. The region Γ , however, defines an algebraic object, namely the multiplicatively closed subset \mathbf{S} of \mathbf{B} . It is therefore natural to expect that this geometric property of Γ is captured by some algebraic property of the ring $\mathbf{S}^{-1}\mathbf{B}$. The next proposition reveals this.

PROPOSITION 3.1.19. *Let Γ be a region in \mathbb{C}^n .*

- (i) *If Γ is codimension k -convex then the following statement holds:*
 (18) *Let \mathbf{p}_f be a prime ideal of height k in $\mathbf{S}^{-1}\mathbf{B}$. Then \mathbf{p}_f is contained in an ideal \mathbf{m}_f of height n .*
 (ii) *Conversely, if Γ is codimension n -convex, then statement (18) implies codimension k -convexity of Γ .*

Proof. (i) Let \mathbf{p} be the contraction of \mathbf{p}_f to \mathbf{B} . Clearly \mathbf{p} is of height k which implies that $V(\mathbf{p})$ is of codimension k . As $\mathbf{p} \cap \mathbf{S} = \emptyset$, by Proposition 3.1.17 $V(\mathbf{p}) \cap \Gamma \neq \emptyset$. So let x be a point in $V(\mathbf{p}) \cap \Gamma$. Then $\mathbf{I}(x)$, the maximal ideal in \mathbf{B} with $V(\mathbf{I}(x)) \supseteq \{x\}$, is of height n containing \mathbf{p} and such that $\mathbf{I}(x) \cap \mathbf{S} = \emptyset$. Let \mathbf{m}_f be the extension in $\mathbf{S}^{-1}\mathbf{B}$ of $\mathbf{I}(x)$. Clearly \mathbf{m}_f is of height n and contains \mathbf{p}_f .

(ii) Now assume that Γ is codimension n -convex. Suppose to the contrary that Γ is not codimension k -convex. Then there exists a prime ideal \mathbf{p} of height k with $V(\mathbf{p}) \cap \Gamma = \emptyset$ and such that for all $f \in \mathbf{p}$, $V(f) \cap \Gamma \neq \emptyset$. This further implies that $\mathbf{p} \cap \mathbf{S} = \emptyset$ and hence that its extension \mathbf{p}_f is a prime ideal in $\mathbf{S}^{-1}\mathbf{B}$. We claim that \mathbf{p}_f is not contained in an ideal of height n . For if \mathbf{m}_f were such a height n ideal containing \mathbf{p}_f then its contraction \mathbf{m} is a maximal ideal in \mathbf{B} with $\mathbf{m} \cap \mathbf{S} = \emptyset$. As \mathbf{m} is of height n and Γ is codimension n -convex, by Proposition 3.1.17 $V(\mathbf{m}) \cap \Gamma \neq \emptyset$. So let $x \in V(\mathbf{m}) \cap \Gamma$. As $V(\mathbf{m})$ is contained $V(\mathbf{p})$, x belongs to $V(\mathbf{p}) \cap \Gamma$ which is a contradiction. \square

Remark. The above proposition has the following important consequence when $\mathbb{K} = \mathbb{C}$. Let Γ be a region in \mathbb{C}^n which is codimension k -convex for all $2 \leq k \leq n$. Then the maximal ideals of $\mathbf{S}^{-1}\mathbb{C}[X_1 \cdots X_n]$ are in one-to-one correspondence with the points of Γ .

We now address the multidimensional stabilization problem considered by Guiver and Bose explained in the beginning of this section.

Here $\mathbf{B} = \mathbb{C}[X_1 \cdots X_n]$, $\Gamma = \bar{U}^n$, the closed unit polydisc and the causal structure is defined by Γ' which is the origin in \mathbb{C}^n . By Theorem 3.1.18 the geometric condition

for stabilizability holds if \bar{U}^n is codimension 2-convex. We now show that in fact \bar{U}^n is codimension k -convex for all $2 \leq k \leq n$.

We make use of the following well-known results (see, for instance, Gunning and Rossi [5]).

T1 (Corollary to Cartan's theorem *B*). Suppose $f_1 \cdots f_k$ are holomorphic functions on a pseudoconvex domain $\Gamma \subset \mathbb{C}^n$ such that the $\{f_i\}$ have no common zeros in Γ . Then there exist holomorphic functions $g_1 \cdots g_k$ such that $\sum f_i g_i = 1$.

T2 (Oka-Weil approximation theorem). Let $\Gamma \subset \mathbb{C}^n$ be a polynomially convex domain. Then any holomorphic function on Γ is uniformly approximable on compact subsets of Γ by polynomials.

PROPOSITION 3.1.20. *The polydisc \bar{U}^n is codimension k -convex for all $2 \leq k \leq n$.*

Proof. Given any variety V with $V \cap \bar{U}^n = \emptyset$, we need to find a polynomial $h \in \mathbb{C}[X_1 \cdots X_n]$ such that $V \subset V(h)$ and $V(h) \cap \bar{U}^n = \emptyset$.

As \bar{U}^n and V are closed subsets (in the \mathbb{C} -topology) there is an open set $W \supset \bar{U}^n$ with $W \cap V = \emptyset$. In fact W can be chosen to be geometrically convex with smooth boundary. Clearly W is also pseudoconvex.

Now let the ideal $\mathbf{I}(V) = (f_1 \cdots f_k)$ be the ideal of the variety V . As $V \cap \bar{U}^n = \emptyset$ the polynomials $\{f_i\}$ do not all vanish at any point in \bar{U}^n . Hence by T1 we can find $g_1 \cdots g_k$, holomorphic on W , such that $\sum f_i g_i = 1$. As W is also polynomially convex, by T2 we can uniformly approximate on the compact set \bar{U}^n , the holomorphic functions $g_1 \cdots g_k$ by polynomials $h_1 \cdots h_k$ so that

$$\sum_{i=1}^k f_i(x) h_i(x) \neq 0 \quad \text{for all } x \in \bar{U}^n.$$

Let $h = \sum f_i h_i$. Clearly $V(h) \cap \bar{U}^n = \emptyset$ and h belongs to $\mathbf{I}(V)$ which implies that $V \subset V(h)$. \square

Thus it follows from the above proposition that the geometric stabilizability condition of Theorem 3.1.18 holds for the n -D stabilization problem described above. Furthermore, if $p = fg^{-1}$ is a stabilizable transfer function with f, g relatively prime in $\mathbb{C}[X_1 \cdots X_n]$ then by the above result there exist h_1, h_2 in $\mathbb{C}[X_1 \cdots X_n]$ such that

$$h_1 f + h_2 g = h \in \mathbf{S}.$$

As h belongs to \mathbf{S} , it is a unit in $\mathbf{A} = \mathbf{S}^{-1}\mathbf{B}$. Hence by the remark following Corollary 2.1.5, a stabilizing controller of p is given by $h_1 h_2^{-1}$.

Consider now the causal structure defined via Γ' , the origin in \mathbb{C}^n ; i.e., let \mathbf{T} be the saturated multiplicatively closed subset of \mathbf{A} consisting of rational functions fg^{-1} such that the variety of f does not contain the origin. By Proposition 2.1.3 all the stabilizing controllers of a strictly causal transfer function are weakly causal. In this case (of n -D systems) we can further conclude that every stabilizable transfer function $p = fg^{-1}$ has a weakly causal controller. This is because if $h_1 f + h_2 g$ is in \mathbf{S} , then by perturbing h_2 (slightly) to h'_2 we can ensure that $h_1 f + h'_2 g$ is still in \mathbf{S} as well as that $h'_2(0) \neq 0$. Then $h_1 h'_2^{-1}$ is in $\mathbf{T}^{-1}\mathbf{A}$ and is therefore a weakly causal controller of p .

In Open Problem 6 in [2] Guiver poses the following question: Given the plant $p = fg^{-1}$ with f, g in $\mathbb{K}[X_1 \cdots X_n]$, is there a stabilizing controller $c = h_1 h_2^{-1}$ with h_1, h_2 also in $\mathbb{K}[X_1 \cdots X_n]$, $\mathbb{K} = \mathbb{R}$ or \mathbb{Q} ? We now answer this question in the affirmative.

THEOREM 3.1.21. *Let $p = fg^{-1}$ be a stabilizable transfer function with f, g relatively prime and in $\mathbb{R}[X_1 \cdots X_n]$. Then there exists a stabilizing controller $c = h_1 h_2^{-1}$ with h_1, h_2 in $\mathbb{Q}[X_1 \cdots X_n]$.*

Proof. As p is stabilizable and as \bar{U}^n is codimension 2-convex the geometric criterion of Theorem 3.1.18 implies that there is a polynomial h belonging to the ideal

$(f, g) \in \mathbb{C}[X_1 \cdots X_n]$ with $V(h) \cap \bar{U}^n = \emptyset$. Consider now the polynomial hh^* in $\mathbb{R}[X_1 \cdots X_n]$ (here h^* denotes the complex conjugate of h). Clearly as $\bar{U}^{n*} = \bar{U}^n$, $V(hh^*) \cap \bar{U}^n = \emptyset$. Equally clearly hh^* belongs to the contraction of the ideal (f, g) to $\mathbb{R}[X_1 \cdots X_n]$. Hence there exist polynomials h_1, h_2 in $\mathbb{R}[X_1 \cdots X_n]$ such that

$$fh_1 + gh_2 = hh^* \in S.$$

As \mathbb{Q} is dense in \mathbb{R} it is possible to perturb h_1 and h_2 to h'_1 and h'_2 in $\mathbb{Q}[X_1 \cdots X_n]$ such that

$$fh'_1 + gh'_2 \in S.$$

Thus, $h'_1 h'^{-1}_2$ is the desired stabilizing controller. \square

Remark. It follows from the above proofs that any compact polynomially convex domain Γ in \mathbb{C}^n is codimension k -convex for all $2 \leq k \leq n$. Hence the geometric stabilizability condition of Theorem 3.1.18 for the case when $\mathbb{K} = \mathbb{C}$ holds for all such Γ . Furthermore, the polydisc \bar{U}^n in Theorem 3.1.21 can be replaced by such Γ which are also symmetric about the real axis.

Note that this is more than what is necessary for the stabilizability problem as it suffices for Γ to be only codimension 2-convex. Thus it is desirable to characterize geometrically such regions. Note that an algebraic characterization of codimension 2-convex regions is available via Proposition 3.1.19.

However, it is *not* possible to replace Γ by any compact region, for instance, by those that are not holomorphically convex (because of Hartog's phenomenon). Thus, even in the 2-D case the above geometric criteria for stabilizability will not hold with the polydisc \bar{U}^2 replaced by a compact annular region. This negatively answers a question raised by Guiver in [2].

3.2. Robustness of stability. We consider here the robustness of stabilizability of n -D systems, i.e., we wish to define a topology on the set of transfer functions \mathbf{F} , the field of fractions of \mathbf{A} (where now $\mathbf{A} = S^{-1}\mathbb{C}[X_1 \cdots X_n]$ and S is the multiplicatively closed subset of \mathbf{B} defined with respect to the polydisc \bar{U}^n) with respect to which Conditions C1 to C4 in § 2.3 are satisfied. This will enable us to carry over the robustness results obtained in the general setting there to the special case of n -D systems here. This is accomplished in a series of steps as follows.

(i) The ring \mathbf{A} here is a topological (in fact, normed) ring with respect to the following norm:

$$\|fg^{-1}\| = \sup_{z \in \bar{U}^n} |f(z)g(z)^{-1}|, \quad fg^{-1} \in \mathbf{A}.$$

Note that this is well defined since fg^{-1} is holomorphic in \bar{U}^n , and is a norm since \bar{U}^n has nonempty interior. It is an easy check that \mathbf{A} is a normed ring under the above norm. Clearly it also follows that if the sequence $f_n g_n^{-1}$ converges to fg^{-1} in this norm topology then the variety $V(f_n)$ converges uniformly to $V(f)$ in \bar{U}^n . \mathbf{A} is Hausdorff being a normed ring; hence C1 is satisfied.

(ii) We now (in the notation of § 2.2) impose a topology τ on $\mathbf{A} \times \mathbf{A}$ such that Condition C2 is satisfied. We describe this topology via a basis of neighborhoods about each (x, y) in $\mathbf{A} \times \mathbf{A}$ as follows:

As \mathbf{A} is a UFD express (x, y) as $h(x', y') = (hx', hy')$ where x', y' in \mathbf{A} are relatively prime. Let $N(h)$, $N(x')$, and $N(y')$ be neighborhoods of h , x' , and y' in the norm topology of \mathbf{A} . Obtain a neighborhood $N(x, y)$ of (x, y) as

$$(19) \quad N(x, y) = \{h''(x'', y'') \mid h'' \in N(h), x'' \in N(x'), y'' \in N(y')\}.$$

The collection $N(x, y)$ obtained this way as $N(h)$, $N(x')$, and $N(y')$ vary over all neighborhoods of h , x' , and y' is a basis for a topology τ on $\mathbf{A} \times \mathbf{A}$. We now show that Condition C2 is satisfied with respect to this topology.

Let (x, y) and (a, b) in $\mathbf{A} \times \mathbf{A}$ satisfy

$$ay = bx.$$

Let $(x, y) = h(x', y')$ and $(a, b) = g(a', b')$, where the pairs (x', y') and (a', b') are each relatively prime. Hence the above equation is equivalent to

$$a'y' = b'x'.$$

Therefore $(a', b') = u(x', y')$ for some (unit) u in \mathbf{A} .

Now let $N(x, y)$ be a given neighborhood of (x, y) specified by neighborhoods $N(h)$, $N(x')$, and $N(y')$ as in (19). (Clearly it suffices to check Condition C2 for such neighborhoods since they form a basis for the topology τ .) Let $N(u)$ be a neighborhood of u . Then

$$N(a', b') = \{u''(x'', y'') \mid u'' \in N(u), x'' \in N(x'), y'' \in N(y')\}$$

is a neighborhood of (a', b') . Let $N(g)$ be a neighborhood of g . Then

$$N(a, b) = \{g''(a'', b'') \mid (a'', b'') \in N(a, b) \text{ and } g'' \in N(g)\}$$

is a neighborhood of (a, b) .

Now let (x'', y'') be in the above given neighborhood $N(x, y)$. Then define

$$(a'', b'') = (gux'', gyy'').$$

Clearly (a'', b'') is in $N(a, b)$ and satisfies $a''y'' = b''x''$. This shows that Condition C2 is satisfied.

(iii) Give $\mathbf{A} \times \mathbf{A}^* \subset \mathbf{A} \times \mathbf{A}$ the subspace topology with respect to τ , also denoted τ . Topologize the set of transfer functions by the quotient topology τ_q via the projection

$$\pi: \mathbf{A} \times \mathbf{A}^* \rightarrow \mathbf{F}.$$

(iv) Clearly from (19) it follows that the product topology τ_p on $\mathbf{A} \times \mathbf{A}$ induced by the norm topology on \mathbf{A} is weaker than the topology τ , i.e., Condition C3 is satisfied.

(v) For any x in \mathbf{A} with $\|x\| < 1$, $1+x$ is a unit in \mathbf{A} . Hence the set of units in \mathbf{A} is open, i.e., C4 is satisfied.

Thus it follows by (i)–(v) above that all the results of § 2.2 on robustness hold in this special case of n -D systems. We therefore have the following theorem.

THEOREM 3.2.22. (1) *Let c be a stabilizing controller of an n -D transfer function p . Then there is a neighborhood of p in the quotient topology τ_q such that every transfer function in this neighborhood is stabilized by the same controller c .*

(2) *Stabilizability is not a generic property of the set of n -D transfer functions.*

Proof. Part (1) follows from the above discussion, i.e., from Theorem 2.3.15.

(2) Consider polynomials $f, g \neq 0$ in $\mathbb{C}[X_1 \cdots X_n]$ such that (a) the singular loci of $V(f)$ and $V(g)$ do not intersect \bar{U}^n ; (b) $V(f)$ and $V(g)$ have a nonempty transversal intersection in \bar{U}^n .

Obviously such f and g exist. Consider the transfer function $p = fg^{-1}$. By the above theory p is not stabilizable. Clearly, there exists a $\delta > 0$ such that for all xx'^{-1} and yy'^{-1} in $S^{-1}\mathbb{C}[X_1 \cdots X_n]$ with

$$|f - xx'^{-1}| < \delta, \quad |g - yy'^{-1}| < \delta,$$

$V(x)$ and $V(y)$ intersect transversally in \bar{U}^n . Then the collection of all such transfer functions $xy'(x'y)^{-1}$ is clearly open in τ_p and therefore in τ_q , and are all non-stabilizable. \square

Remark. We emphasize here once again that the topology τ on $\mathbf{A} \times \mathbf{A}$ is not the product topology τ_P . In fact the robustness result is not valid with respect to τ_P . Actually even more is true, viz., with respect to τ_P every neighborhood of a stabilizable transfer function contains nonstabilizable ones as the following simple argument demonstrates.

Let $p = fg^{-1}$, where f and g are relatively prime be stabilizable, i.e., $V(f)$ and $V(g)$ do not intersect in \bar{U}^n . In the product topology every neighborhood of p will contain transfer functions of the kind $hf(kg)^{-1}$, where hf and kg are relatively prime but where $V(h)$ and $V(k)$ intersect in \bar{U}^n . Such plants are clearly not stabilizable.

Remark. From the above proof it is clear that every neighborhood of an n -D transfer function $p = fg^{-1}$, with f, g relatively prime and such that $V(f) \cap V(g)$ contains points of the boundary of \bar{U}^n but not its interior U^n , contains stabilizable transfer functions. Thus such transfer functions belong to the boundary of the closed subset of nonstabilizable transfer functions and can therefore be perturbed and made stabilizable. A more detailed study of this boundary will appear elsewhere.

Acknowledgments. We are grateful to G. Misra, A. Parameswaran, S. Patkar, and A. R. Shastri for many useful discussions. We would also like to thank a very knowledgeable referee for his comments which led to improvements in the presentation of this paper.

REFERENCES

- [1] M. F. ATIYAH AND I. G. McDONALD, *Introduction to Commutative Algebra*, Addison-Wesley, Reading, MA, 1969.
- [2] N. K. BOSE, ED., *Multidimensional Systems Theory*, D. Reidel, Boston, MA, 1985.
- [3] C. A. DESOER AND C. L. GUSTAFSON, *Algebraic theory of linear multivariable feedback systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 909-917.
- [4] C. A. DESOER, R. W. LIU, J. MURRAY, AND R. SAEKS, *Feedback system design: the fractional representation approach to analysis and synthesis*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 399-412.
- [5] R. GUNNING AND H. ROSSI, *Analytic Functions in Several Complex Variables*, Prentice-Hall, Englewood Cliffs, NJ, 1965.
- [6] R. HARTSHORNE, *Algebraic Geometry*, Springer-Verlag, Berlin, New York, 1977.
- [7] V. R. SULE, Ph.D. thesis, Dept of Elect. Engrg., I.I.T., Bombay, 1990.
- [8] M. VIDYASAGAR, *The graph metric for unstable plants and robustness estimates for feedback stability*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 403-418.
- [9] ———, *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.
- [10] M. VIDYASAGAR, H. SCHNIDER, AND B. A. FRANCIS, *Algebraic and topological aspects of feedback stabilization*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 880-894.
- [11] S. Q. ZHU, *Graph topology and gap topology for unstable systems*, IEEE Trans. Automat. Control., AC-34 (1989), pp. 848-855.

UNBOUNDED SOLUTIONS TO THE LINEAR QUADRATIC CONTROL PROBLEM*

G. DA PRATO[†] AND M. C. DELFOUR[‡]

Abstract. Examples are presented to show that the solution of the operational algebraic Riccati equation can be an unbounded operator for infinite dimensional systems in a Hilbert space even with bounded control and observation operators. This phenomenon is connected to the presence of a continuous spectrum in one of the operators. The object of this paper is to fill up the gap in the classical linear quadratic theory. The key step is the introduction of the set of stabilizable initial conditions. Then a new simple approach to the linear-quadratic problem is presented that provides the connection with the notion of approximate stabilizability for the triplet (A, B, C) .

Key words. linear quadratic, stabilizability, Riccati equation

AMS(MOS) subject classifications. 49A22

1. Introduction. The infinite time, linear quadratic, optimal control theory for infinite-dimensional systems in Hilbert spaces with bounded control and observation operators has been extensively studied (see, for instance, the book by Curtain and Pritchard [1]). Typically, let H (*state space*), U (*control space*), and Y (*observation space*) be three Hilbert spaces. Let $A: D(A) \subset H \rightarrow H$ be the infinitesimal generator of a strongly continuous semigroup e^{tA} and let $B: U \rightarrow H$ and $C: H \rightarrow Y$ be continuous linear operators. The state $x(t)$ at time $t \geq 0$ is given by

$$(1.1) \quad x(t) = e^{tA}h + \int_0^t e^{(t-s)A}Bu(s) ds, \quad t \geq 0,$$

and the cost function by

$$(1.2) \quad J(u, h) = \int_0^\infty \{|Cx(s)|^2 + |u(s)|^2\} ds.$$

Under the standard (A, B, C) stabilizability hypothesis for the triplet (A, B, C) ,

$$(1.3) \quad \forall h \in H, \quad \exists u \in L^2(0, \infty; U) \quad \text{such that } J(u, h) < \infty,$$

it is well known that the corresponding algebraic operator Riccati equation

$$(1.4) \quad A^*P + PA - PBB^*P + C^*C = 0$$

has a minimum positive symmetrical bounded solution \underline{P} ; that is,

$$(1.5) \quad \underline{P}: H \rightarrow H \text{ is linear and continuous (bounded),}$$

$$(1.6) \quad \underline{P}^* = \underline{P} \text{ (symmetry), } \forall h \in H, (\underline{P}h, h) \geq 0 \text{ (positivity),}$$

and for any other solution of (1.4) verifying (1.5) and (1.6)

$$(1.7) \quad \forall h \in H, \quad (Qh, h) \geq (\underline{P}h, h) \text{ (minimality).}$$

* Received by the editors January 31, 1990; accepted for publication (in revised form) November 20, 1990.

[†] Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, Italy. The author has been partially supported by the Italian National Project M.P.I. Equazioni di Evoluzione e Applicazioni Fisico-Matematiche.

[‡] Centre de Recherches Mathématiques et Département de Mathématiques et de Statistique, Université de Montréal, C.P. 6128 A, Montréal, Québec, Canada H3C 3J7. The author has been supported by a Killam fellowship from Canada Council, Natural Sciences and Engineering Research Council of Canada operating grant A-8730, and a Fonds pour la Formation de Chercheurs et l'Aide à la Recherche (le Fonds FCAR, Québec.) grant from the Ministère de l'Éducation du Québec.

The authors have recently constructed examples where the system is not stabilizable, and yet the algebraic Riccati equation has a positive selfadjoint unbounded solution (cf. [2]). This phenomenon is intimately related to the fact that only a dense subset of initial conditions are (A, B, C) stabilizable. This has many interesting implications for infinite-dimensional control systems. For instance, it points out that definitions of stabilizability (here, (A, B, I) stabilizability) that assume the existence of a bounded feedback operator really contain two hypotheses in one: the existence of a feedback operator that stabilizes all initial conditions in H , and the boundedness of this operator. Example 6.1 in § 6 describes a control system that can only be stabilized by an unbounded feedback operator for all initial conditions in H .

The object of this paper is to fill the gap in the theory. Under no stabilizability hypothesis, we a priori define the set Σ of initial states that can be (A, B, C) stabilized and show that it can be given a natural Hilbert space structure. When Σ is dense in the space of initial conditions, we construct the smallest or minimum positive self-adjoint unbounded solution to the algebraic Riccati equation. A new technique is introduced to directly obtain the semigroup associated with the closed loop system and the properties of the feedback operator. If the usual detectability hypothesis is added, we recover that the closed loop system is exponentially stable. Examples are also included to illustrate the theoretical considerations. Extensions to systems with unbounded control and observation operators are possible and will be reported in a forthcoming paper. We felt that it was more instructive to first illustrate the phenomenon and the main features of the theory for the bounded case.

Notation. The space of continuous linear operators from a Hilbert space X to another Hilbert space Y will be denoted by $\mathcal{L}(X; Y)$. When $X = Y$, the cone of continuous linear operators in $\mathcal{L}(X; X)$ verifying conditions (1.5) and (1.6) will be denoted $\Sigma^+(X)$. \mathbf{R} will be the field of all real numbers and \mathbf{N} the set of integers greater than or equal to 1.

2. Problem formulation. Let H, U, Y, A, B , and C be as defined in § 1. Consider the mild solution of the system

$$(2.1) \quad \begin{aligned} x'(s) &= Ax(s) + Bu(s), \quad s \geq 0, \\ x(0) &= h, \end{aligned}$$

and the associated cost function

$$(2.2) \quad J(u, h) = \int_0^\infty \{|Cx(s)|^2 + |u(s)|^2\} ds.$$

A *mild solution* of (2.1) is a continuous function $x: [0, \infty[\rightarrow H$ verifying (1.1). Denote by V the value function

$$(2.3) \quad V(h) = \inf \{J(u, h): u \in L^2(0, \infty; U)\}$$

with domain

$$(2.4) \quad \text{dom } V = \{h \in H: V(h) < \infty\},$$

which will be referred to as the *domain of stabilizability* for the triple (A, B, C) . Observe that under the (A, B, C) stabilizability condition (1.3) $\text{dom } V = H$.

3. An example of unbounded solution to the Riccati equation. Let $H = \ell^2$ be the Hilbert space of all sequences $x = \{x_n\}_{n \in \mathbf{N}}$, with norm

$$(3.1) \quad |x|^2 = \sum_{k=1}^{\infty} x_k^2.$$

Let $\{e_k\}$ be the orthonormal basis in ℓ^2

$$(3.2) \quad (e_k)_n = \delta_{kn}, \quad k \in \mathbf{N}.$$

Define the bounded operators

$$(3.3) \quad Ae_k = \frac{k}{k+1} e_k, \quad Be_k = \frac{\sqrt{2k+1}}{k+1} e_k, \quad k \in \mathbf{N}.$$

Note that their spectra are made up of a point and a continuous part

$$(3.4) \quad \sigma_p(A) = \left\{ \frac{k}{k+1} : k \in \mathbf{N} \right\}, \quad \sigma_c(A) = \{1\},$$

$$(3.5) \quad \sigma_p(B) = \left\{ \frac{\sqrt{2k+1}}{k+1} : k \in \mathbf{N} \right\}, \quad \sigma_c(B) = \{0\}.$$

Associate with A and B the control system

$$(3.6) \quad \begin{aligned} x'(s) &= Ax(s) + Bu(s), \quad s \geq 0, \\ x(0) &= h, \end{aligned}$$

and the cost function

$$(3.7) \quad J(u, h) = \int_0^\infty \{|x(s)|^2 + |u(s)|^2\} ds.$$

Here the observation operator C is the identity operator on H . If the pair (A, B) was stabilizable, there would exist a symmetric positive bounded linear operator P_∞ on H that would be the minimum solution of the algebraic operator Riccati equation

$$(3.8) \quad P_\infty A + A^* P_\infty - P_\infty B B^* P_\infty + I = 0,$$

in the sense of conditions (1.5) and (1.6). Here A , B , and I are diagonal operators, and it is easy to check that the only positive selfadjoint solution to (3.8) is the diagonal unbounded operator

$$(3.9) \quad P_\infty e_k = (k+1) e_k, \quad k \in \mathbf{N}.$$

This means that only initial conditions h in the domain $D(P_\infty^{1/2})$ of $P_\infty^{1/2}$

$$(3.10) \quad \begin{aligned} D(P_\infty^{1/2}) &= \left\{ x \in \ell^2 : \sum_{k=1}^\infty (k+1) x_k^2 < \infty \right\}, \\ P_\infty^{1/2} e_k &= \sqrt{k+1} e_k \end{aligned}$$

can be stabilized, and that for all others

$$(3.11) \quad J(u, h) = \infty, \quad h \notin D(P_\infty^{1/2}).$$

Hence $\text{dom } V = D(P_\infty^{1/2})$ in this example.

The interpretation of this phenomenon is that, for initial conditions $h \notin D(P_\infty^{1/2})$, the corresponding state x cannot be stabilized with a finite energy control u in $L^2(0, \infty; H)$. Yet the closed loop system is given by the operator

$$(3.12) \quad A - B B^* P_\infty = -I,$$

which is exponentially stable in H , and for all h in H the solution x^* of the closed loop system

$$(3.13) \quad \begin{aligned} x'(s) &= [A - B B^* P_\infty] x(s), \quad s \geq 0, \\ x(0) &= h, \end{aligned}$$

is given by $x^*(s) = e^{-s}h$, which belongs to $L^2(0, \infty; H)$, whereas the optimal control u^* is given by

$$(3.14) \quad u^*(s) = -B^*P_\infty x^*(s) = -B^*P_\infty e^{-s}h.$$

So u belongs to $L^2(0, \infty; H)$ if and only if $h \in D(P_\infty^{1/2})$.

Finally, it is useful to note that

$$(3.15) \quad B^*P_\infty e_k = \sqrt{2k+1} e_k, \quad D(B^*P_\infty) = D(P_\infty^{1/2})$$

and that

$$(3.16) \quad BB^*P_\infty e_k = \frac{2k+1}{k+1} e_k, \quad D(BB^*P_\infty) = H.$$

4. Asymptotic behaviour of the solution $P(t)$ to the associated differential operator Riccati equation. It is well known that for any fixed $T > 0$, we can associate with the control problem (2.1)–(2.2) the mild solution $P \in C_s([0, \infty[; \Sigma^+(H))$ of the differential operator Riccati equation

$$(4.1) \quad \begin{aligned} P'(t) &= A^*P(t) + P(t)A - P(t)BB^*P(t) + C^*C \quad \text{in } [0, T], \\ P(0) &= 0. \end{aligned}$$

We say that P in $C_s([0, T]; \Sigma^+(H))$ is a *mild solution* of the Riccati differential equation (4.1) if P verifies the integral equation

$$P(t)x = \int_0^t \{e^{(t-s)A^*} [C^*C - P(s)BB^*P(s)] e^{(t-s)A} x\} ds$$

for all x in H (for example, see Curtain and Pritchard [1] for basic results on existence and uniqueness). We have denoted by $C_s([0, \infty[; \Sigma^+(H))$ the set of all mappings $T: [0, \infty[\rightarrow \Sigma^+(H)$, such that $T(\cdot)x$ is continuous for all $x \in H$.

For each $h \in H$ the function $(P(\cdot)h, h)$ is nondecreasing. Moreover, the following identity holds:

$$(4.2) \quad (P(t)h, h) + \int_0^t |u(s) + B^*P(t-s)x(s)|^2 ds = \int_0^t \{|Cx(s)|^2 + |u(s)|^2\} ds, \quad \forall u \in L_{loc}^2(0, \infty; U).$$

To obtain identity (4.2) fix $t > 0$, multiply both sides of (4.1) evaluated at $t-s$, $t \geq s \geq 0$, by $x(s)$, use (2.1) to eliminate $x'(s)$, and integrate with respect to s from 0 to t .

Define the function

$$(4.3) \quad h \rightarrow \phi(h) = \lim_{t \rightarrow \infty} (P(t)h, h): H \rightarrow [0, \infty].$$

The function ϕ is convex, proper,¹ and lower semicontinuous with domain

$$(4.4) \quad \Sigma = \{h \in H: \phi(h) < \infty\}.$$

LEMMA 4.1. *The following properties are verified:*

- (i) *For all h and k in Σ , $(P(\cdot)h, k)$ is bounded;*
- (ii) *Σ is a vector subspace of H ;*
- (iii) *For all h and k in Σ , the following limit exists*

$$(4.5) \quad \psi(h, k) = \lim_{t \rightarrow \infty} (P(t)h, k).$$

¹ A convex function $f: H \rightarrow [0, \infty]$ is said to be proper if $f(x) < \infty$ for at least one x and $f(x) > -\infty$ for every x (cf. R. T. Rockafellar [7, p. 24]).

Moreover, ψ is a bilinear form on $\Sigma \times \Sigma$ and

$$(4.6) \quad \psi(h, h) = \phi(h), \quad \forall h \in H.$$

Proof. (i) For all h and k in Σ and $t \geq 0$, we have

$$(4.7) \quad |(P(t)h, k)|^2 \leq (P(t)h, h)(P(t)k, k) \leq \phi(h)\phi(k),$$

and the conclusion follows.

(ii) For all h in Σ and λ in \mathbf{R} , $\phi(\lambda h) = \lambda^2 \phi(h)$, and hence $\lambda h \in \Sigma$. For all h and k in Σ

$$(P(t)(h+k), h+k) = (P(t)h, h) + (P(t)k, k) + 2(P(t)h, k)$$

and from (i), $\phi(h+k) \leq [\phi(h)^{1/2} + \phi(k)^{1/2}]^2$. Thus Σ is a linear subspace of H . Part (iii) is an immediate consequence of parts (i) and (ii). \square

Define the following inner product on Σ

$$(4.8) \quad (h, k)_\Sigma = (h, k) + \psi(h, k),$$

which makes it a pre-Hilbert space.

LEMMA 4.2. *The space Σ endowed with the inner product (4.8) is a Hilbert space.*

Proof. It is sufficient to show that Σ is complete with respect to the norm

$$(4.9) \quad |h|_\Sigma = [|h|^2 + \phi(h)]^{1/2}.$$

Let $\{h_n\}$ be a Cauchy sequence in Σ . Then there exists $h \in H$ such that $h_n \rightarrow h$. Moreover, there exists $\lambda \geq 0$ such that

$$(4.10) \quad |h_n|^2 + \phi(h_n) \rightarrow \lambda$$

and

$$(4.11) \quad \phi(h_n) \rightarrow \lambda - |h|^2.$$

By lower semicontinuity of ϕ , we have

$$(4.12) \quad \lambda - |h|^2 = \lim_{n \rightarrow \infty} \phi(h_n) \geq \phi(h),$$

and, by definition of Σ , h belongs to Σ . Finally, for each $\varepsilon > 0$, there exists a positive integer $N(\varepsilon)$ such that

$$|h_n - h_m|_\Sigma^2 = |h_n - h_m|^2 + \phi(h_n - h_m) \leq \varepsilon, \quad \forall m, n \geq N(\varepsilon).$$

As n goes to infinity, we get

$$|h - h_m|^2 + \phi(h - h_m) \leq \varepsilon, \quad \forall m \geq N(\varepsilon),$$

by continuity of the norm in H and lower semicontinuity of ϕ . This shows that $h_n \rightarrow h$ in Σ and completes the proof. \square

We have constructed the space Σ of initial conditions for which the expression $(P(t)h, h)$ has a limit. In general, its closure in H will not be dense, and it will be natural to decompose H as a direct sum

$$(4.13) \quad H = \bar{\Sigma} \oplus \Sigma^\perp,$$

where $\bar{\Sigma}$ is the closure of Σ in H , and Σ^\perp is the orthogonal complement to Σ in H . In the following, we identify the elements of the dual H' of H with those of H . We denote by Σ' the dual of Σ .

PROPOSITION 4.3. Assume that Σ is dense in H . Then there exists a unique linear operator $P_\infty \in \mathcal{L}(\Sigma; \Sigma')$ such that

$$(4.14) \quad \langle P_\infty h, k \rangle_\Sigma = \psi(h, k), \quad \forall h, k \in \Sigma.$$

P_∞ can also be viewed as a closed selfadjoint positive operator on H with dense domain

$$(4.15) \quad D(P_\infty) = \{h \in \Sigma: \psi(h, \cdot) \text{ is continuous in } H\}.$$

We have

$$(4.16) \quad \phi(h) = (P_\infty h, h), \quad \forall h \in D(P_\infty),$$

$$(4.17) \quad \psi(h, k) = (P_\infty h, k), \quad \forall h \in D(P_\infty), \forall k \in H,$$

and the subdifferential of ϕ is given by

$$(4.18) \quad \frac{1}{2}\partial\phi(h) = \begin{cases} P_\infty h, & \text{if } h \in D(P_\infty), \\ \emptyset, & \text{if } h \notin D(P_\infty); \end{cases}$$

that is,

$$\partial\phi(h) = \{p \in H \mid \forall v \in D(P_\infty), \langle p, v \rangle \leq d\phi(h; v)\},$$

where $d\phi(h; v)$ is the Gâteaux semiderivative at h in the direction v .

Moreover, $P_\infty^{1/2}$ is well defined and

$$(4.19) \quad D(P_\infty^{1/2}) = \Sigma = [D(P_\infty), H]_{1/2},$$

where $[X, Y]_{1/2}$ denotes the interpolation space between Y and its dense subspace X (see Lions and Peetre [6] or Lions and Magenes [5] for the theory of interpolation spaces).

Proof. By definition of the inner product on Σ , the symmetrical bilinear form ψ on $\Sigma \times \Sigma$ is continuous, and there exists a unique $P_\infty \in \mathcal{L}(\Sigma; \Sigma')$ such that (4.14) is verified. Moreover, ψ is Σ - H coercive and P_∞ is a self-adjoint operator in H with domain $D(P_\infty)$. Expression (4.18) follows from the fact that ϕ is lower semicontinuous. Hence $\partial\phi(\cdot)$ is maximal monotone on H as a set-valued function. Finally, the positive self-adjoint operator P_∞ has a positive square root $P_\infty^{1/2}$, which is a closed linear operator on H with dense domain $D(P_\infty^{1/2})$, which coincides with Σ . \square

Assume now that Σ is not dense in H , and denote by $\bar{\Sigma}$ the closure of Σ in H . Then we have the following similar result.

COROLLARY. There exists a unique linear operator $P_\infty \in \mathcal{L}(\Sigma; \Sigma')$ such that

$$(4.20) \quad \langle P_\infty h, k \rangle_\Sigma = \psi(h, k), \quad \forall h, k \in \Sigma.$$

P_∞ can also be viewed as a closed selfadjoint positive operator on $\bar{\Sigma}$ with dense domain

$$(4.21) \quad D(P_\infty) = \{h \in \Sigma: \psi(h, \cdot) \text{ is continuous in } \bar{\Sigma}\}.$$

We have

$$(4.22) \quad \phi(h) = (P_\infty h, h), \quad \forall h \in D(P_\infty),$$

$$(4.23) \quad \psi(h, k) = (P_\infty h, k), \quad \forall h \in D(P_\infty), \quad \forall k \in \bar{\Sigma},$$

and the subdifferential of ϕ is given by

$$(4.24) \quad \frac{1}{2}\partial\phi(h) = \begin{cases} P_\infty h, & \text{if } h \in D(P_\infty), \\ \emptyset, & \text{if } h \notin D(P_\infty). \end{cases}$$

Moreover, $P_\infty^{1/2}$ is well defined and

$$(4.25) \quad D(P_\infty^{1/2}) = \Sigma = [D(P_\infty), \bar{\Sigma}]_{1/2}. \quad \square$$

5. Existence of the optimal control and optimal closed loop system. In this section we use the asymptotic properties obtained in § 4 to solve the optimal control problem (2.1)–(2.2). In addition, we study the mapping between the initial conditions and the optimal state and control.

THEOREM 5.1. *The following statements hold:*

(i) *Given any h in H , either $h \notin \Sigma$ and*

$$(5.1) \quad J(u, h) = +\infty, \quad \forall u \in L^2(0, \infty; U) \quad \text{and} \quad V(h) = \phi(h) = +\infty,$$

or $h \in \Sigma$ and there exists a unique optimal control $\hat{u}(\cdot, h)$ in $L^2(0, \infty; U)$ such that

$$(5.2) \quad J(\hat{u}(\cdot, h), h) = V(h) = \phi(h).$$

(ii) *The mapping*

$$(5.3) \quad \Sigma \rightarrow L^2(0, \infty; U), \quad h \rightarrow \hat{u}(\cdot, h)$$

is linear and continuous.

(iii) *Denote by $\hat{x}(\cdot, h)$ the optimal state corresponding to the optimal control $\hat{u}(\cdot, h)$ and set*

$$(5.4) \quad S_\Sigma(t)h = \hat{x}(t, h), \quad t \geq 0, \quad h \in \Sigma.$$

Then $S_\Sigma(\cdot)$ is a strongly continuous semigroup in Σ .

(iv) *Let A_Σ be the infinitesimal generator of $S_\Sigma(\cdot)$. For all $h \in D(A_\Sigma)$, we have that*

$$(5.5) \quad \hat{u}(\cdot, h) \in H^1(0, \infty, U) \quad \text{and} \quad \hat{u}'(\cdot, h) = \hat{u}(\cdot, A_\Sigma h),$$

$$(5.6) \quad C\hat{x}(\cdot, h) \in H^1(0, \infty; Y), \quad \hat{x}'(\cdot, h) = \hat{x}(\cdot, A_\Sigma h), \quad \text{and} \quad D(A_\Sigma) \subset D(A).$$

(v) *For all h in $D(A_\Sigma)$ the map*

$$(5.7) \quad h \rightarrow \hat{u}(0, h) : D(A_\Sigma) \rightarrow U$$

is linear and continuous. Its closure in Σ generates an unbounded linear operator

$$(5.8) \quad K : D(K) \subset \Sigma \rightarrow U \quad \text{such that} \quad D(A_\Sigma) \subset D(K)$$

and

$$(5.9) \quad D(A_\Sigma) = D(A) \cap D(K), \quad A_\Sigma h = Ah + BKh.$$

Moreover, for all h in $D(A_\Sigma)$ and $t \in [0, \infty[$, $\hat{x}(t, h) \in D(A_\Sigma)$,

$$(5.10) \quad A_\Sigma \hat{x}(t, h) = A\hat{x}(t, h) + B\hat{u}(t, h) = [A + BK]\hat{x}(t, h),$$

$$(5.11) \quad \hat{u}(t, h) = K\hat{x}(t, h).$$

(vi) *For all h in $D(A_\Sigma)$,*

$$(5.12) \quad Kh = \lim_{t \rightarrow \infty} -B^*P(t)h,$$

and for all h in Σ and almost all t in $[0, \infty[$

$$\hat{u}(t, h) = K\hat{x}(t, h), \quad \hat{x}(t, h) \in D(K).$$

*When $\bar{\Sigma} = H$ the closure K_∞ of the operator $-B^*P_\infty$ in Σ coincides with K on $D(A_\Sigma)$.*

Proof. (i). By definition of Σ , for all $h \notin \Sigma$ $\lim_{t \rightarrow \infty} (P(t)h, h) = \infty$ and, in view of identity (4.2),

$$(P(t)h, h) \leq J(u, h), \quad \forall u \in L^2(0, \infty; U), \quad \forall t \geq 0.$$

By letting t go to infinity, we obtain (5.1). When $h \in \Sigma$, identity (4.2) yields

$$\phi(h) \leq J(u, h), \quad \forall u \in L^2(0, \infty; U).$$

For each $t > 0$, let (x_t, u_t) be defined by

$$(5.13) \quad \begin{aligned} x_t'(s) &= Ax_t(s) - BB^*P(t-s)x_t(s), \quad \text{in } [0, t], & x_t(0) &= h, \\ u_t(s) &= -B^*P(t-s)x_t(s), \quad \text{in } [0, t]. \end{aligned}$$

The pair (x_t, u_t) is the optimal solution on the interval $[0, t]$. Consider the extension \hat{u}_t of u_t from $[0, t]$ to $[0, \infty[$

$$(5.14) \quad \hat{u}_t(s) = \begin{cases} u_t(s), & \text{if } 0 \leq s \leq t, \\ 0, & \text{if } s > t, \end{cases}$$

and let \hat{x}_t be the corresponding extension of the solution x_t of the state equation on $[0, t]$

$$\hat{x}_t(s) = \begin{cases} x_t(s), & \text{if } 0 \leq s \leq t, \\ 0, & \text{if } s > t. \end{cases}$$

Again by (4.2) and (5.14)

$$(5.15) \quad (P(t)h, h) = \int_0^t \{|Cx_t(s)|^2 + |u_t(s)|^2\} ds \geq \int_0^\infty |\hat{u}_t(s)|^2 ds.$$

Hence for any sequence $\{t_n\}$, $t_n \rightarrow \infty$, the sequence $\{\hat{u}_{t_n}\}$ is bounded in $L^2(0, \infty; U)$. So there exists \hat{u} in $L^2(0, \infty; U)$ and a subsequence of $\{t_n\}$ (still denoted $\{t_n\}$) such that

$$(5.16) \quad \hat{u}_{t_n} \rightarrow \hat{u} \quad \text{in } L^2(0, \infty; U)\text{-weak}.$$

Denote by \hat{x} the solution of

$$(5.17) \quad \hat{x}'(s) = A\hat{x}(s) + B\hat{u}(s), \quad \text{for } s \geq 0, \quad \hat{x}(0) = h.$$

Then for any fixed $T > 0$ and $t_n > T$

$$\hat{u}_{t_n} \rightarrow \hat{u} \quad \text{in } L^2(0, T; U)\text{-weak}, \quad \hat{x}_{t_n} \rightarrow \hat{x} \quad \text{in } L^2(0, T; H)\text{-weak}.$$

For $t_n > T$, however,

$$(P(t_n)h, h) \geq \int_0^T \{|Cx_{t_n}(s)|^2 + |u_{t_n}(s)|^2\} ds$$

and by weak lower semicontinuity

$$\phi(h) \geq \int_0^T \{|C\hat{x}(s)|^2 + |\hat{u}(s)|^2\} ds.$$

As T goes to infinity

$$(5.18) \quad \phi(h) \geq \int_0^\infty \{|C\hat{x}(s)|^2 + |\hat{u}(s)|^2\} ds = J(\hat{u}, h).$$

Combining (5.18) and (5.13) it follows that there exists $\hat{u} = \hat{u}(\cdot, h) \in L^2(0, \infty; U)$ such that

$$J(\hat{u}, h) \leq \phi(h) \leq J(u, h), \quad \forall u \in L^2(0, \infty; U).$$

It follows that $V(h) \leq J(\hat{u}, h) \leq \phi(h) \leq V(h)$. This establishes (5.2). As for the uniqueness of \hat{u} , assume that \hat{u}_1 and \hat{u}_2 are two optimal controls in $L^2(0, \infty; U)$. Then $J(\hat{u}_1, h) = J(\hat{u}_2, h) = V(h)$. So for $\hat{u}_1 \neq \hat{u}_2$

$$\begin{aligned} J((\hat{u}_1 + \hat{u}_2)/2, h) &= \frac{1}{2}[J(\hat{u}_1, h) + J(\hat{u}_2, h)] - J((\hat{u}_1 - \hat{u}_2)/2, h) \\ &= V(h) - J((\hat{u}_1 - \hat{u}_2)/2, h) \leq V(h) - \frac{1}{4}\|\hat{u}_1 - \hat{u}_2\|^2 < V(h), \end{aligned}$$

which contradicts the optimality of \hat{u}_1 and \hat{u}_2 .

(ii) Let \hat{u}_t be defined by (5.13), then

$$(5.19) \quad \|\hat{u}_t\|_{L^2(0, \infty; U)}^2 \leq (P(t)h, h) \leq |h|_{\Sigma}^2.$$

Moreover, since the optimal control is unique, we have proved in part (i) that

$$\lim_{t \rightarrow \infty} \hat{u}_t = \hat{u} \quad \text{in } L^2(0, \infty; U)\text{-weak, for any } h \in \Sigma.$$

We now prove that $\hat{u}_t \rightarrow \hat{u}$ in $L^2(0, \infty; U)$ -strong. By optimality of the pair (x_t, u_t) on $[0, t]$

$$J'(u_t, h) = \inf \{J'(v, h) : v \in L^2(0, \infty; U)\},$$

where

$$J'(v, h) = \int_0^t \{|Cx(s; v)|^2 + |v(s)|^2\} ds.$$

We want to prove that $\lim_{t \rightarrow \infty} J'(u_t, h) = J(\hat{u}, h)$. By definition of the minimizing element u_t on $[0, t]$

$$J'(u_t, h) \leq J'(\hat{u}(\cdot, h), h) = \int_0^t \{|C\hat{x}(s, \hat{u}(\cdot, h))|^2 + |\hat{u}(s, h)|^2\} ds$$

and necessarily

$$\limsup_{t \rightarrow \infty} J'(u_t, h) \leq \int_0^\infty \{|C\hat{x}(s, \hat{u}(\cdot, h))|^2 + |\hat{u}(s, h)|^2\} ds = J(\hat{u}(\cdot, h), h).$$

We have shown in (i) that $\hat{u}_t \rightarrow \hat{u}$ in $L^2(0, \infty; U)$ -weak, and we can show by the same technique that $\{C\hat{x}_t\}$ is bounded in $L^2(0, \infty; Y)$, and that weak subsequences $\{C\hat{x}_{t_n}\}$ converging to some y in $L^2(0, \infty; Y)$ can be extracted as follows:

$$C\hat{x}_{t_n} \rightarrow y, \quad \text{in } L^2(0, \infty; Y)\text{-weak.}$$

By continuity of the state $x(\cdot; u)$ with respect to the control u on a finite time interval $[0, T]$, $T > 0$, the map $u \rightarrow x(\cdot; u) : L^2(0, T; U) \rightarrow L^2(0, T; H)$ is weakly continuous and, finally,

$$u \rightarrow Cx(\cdot; u) : L^2(0, T; U) \rightarrow L^2(0, T; Y)$$

is also weakly continuous. This implies that for all $T > 0$, $y = C\hat{x}(\hat{u}, h)$ in $L^2(0, T; Y)$ and hence in $L^2(0, \infty; Y)$. As a result,

$$\hat{u}_t \rightarrow \hat{u}, \quad \text{in } L^2(0, \infty; U)\text{-weak} \quad \text{and} \quad C\hat{x}_t \rightarrow C\hat{x}, \quad \text{in } L^2(0, \infty; Y)\text{-weak.}$$

The functional

$$(v, y) \rightarrow \int_0^\infty \{|y(s)|^2 + |v(s)|^2\} ds : L^2(0, \infty; U) \times L^2(0, \infty; Y) \rightarrow \mathbf{R}$$

is, however, weakly lower semicontinuous and necessarily

$$\liminf_{t \rightarrow \infty} \int_0^\infty [|C\hat{x}_t|^2 + |\hat{u}_t|^2] ds \geq \int_0^\infty [|C\hat{x}|^2 + |\hat{u}|^2] ds;$$

that is,

$$\liminf_{t \rightarrow \infty} J'(u_t, h) \geq J(\hat{u}, h).$$

Finally,

$$J(\hat{u}, h) \leq \liminf_{t \rightarrow \infty} J'(u_t, h) \leq \limsup_{t \rightarrow \infty} J'(u_t, h) \leq J(\hat{u}, h),$$

and this proves that $\lim_{t \rightarrow \infty} J'(u_t, h) = J(\hat{u}, h)$.

The strong continuity will now be obtained by the following simple computation:

$$\begin{aligned} \|C\hat{x}_t - C\hat{x}\|^2 + \|\hat{u}_t - u\|^2 &= \|C\hat{x}_t\|^2 + \|\hat{u}_t\|^2 + \|C\hat{x}\|^2 + \|\hat{u}\|^2 - 2(C\hat{x}_t, C\hat{x}) - 2(\hat{u}_t, \hat{u}) \\ &= J'(u_t, h) + J(\hat{u}, h) - 2(C\hat{x}_t, C\hat{x}) - 2(\hat{u}_t, \hat{u}). \end{aligned}$$

As t goes to ∞ , $J'(u_t, h) \rightarrow J(u, h)$ and, by weak convergence,

$$(C\hat{x}_t, C\hat{x}) \rightarrow (C\hat{x}, C\hat{x}) = \|C\hat{x}\|^2 \quad \text{and} \quad (\hat{u}_t, \hat{u}) \rightarrow (\hat{u}, \hat{u}) = \|\hat{u}\|^2.$$

So we conclude that

$$\lim_{t \rightarrow \infty} \{\|C\hat{x}_t - C\hat{x}\|^2 + \|\hat{u}_t - u\|^2\} = 2J(\hat{u}, h) - 2[\|C\hat{x}\|^2 + \|\hat{u}\|^2] = 0$$

and that

$$\hat{u}_t \rightarrow \hat{u}, \quad \text{in } L^2(0, \infty; U)\text{-strong} \quad \text{and} \quad C\hat{x}_t \rightarrow C\hat{x}, \quad \text{in } L^2(0, \infty; Y)\text{-strong}.$$

By (5.18) and by the uniform boundedness theorem, it follows that the mapping $h \rightarrow \hat{u}(\cdot, h): \Sigma \rightarrow L^2(0, \infty; U)$ is linear and continuous.

(iii) First, note that, by Bellman's optimality principle, we have $\hat{x}(t, h) \in \Sigma$ for all $h \in \Sigma$ and

$$(5.20) \quad \hat{x}(t+s, h) = \hat{x}(t; \hat{x}(s, h)), \quad \forall t \geq 0, \quad \forall s \geq 0,$$

$$(5.21) \quad V(\hat{x}(t, h)) = \int_t^\infty \{|C\hat{x}(s, h)|^2 + |\hat{u}(s, h)|^2\} ds.$$

Thus $S_\Sigma(t)$ is a linear operator in Σ for all $t \geq 0$. We prove now that $S_\Sigma(t)$ is bounded in Σ . By (5.17) we have

$$(5.22) \quad \hat{x}(t, h) = e^{tA}h + \int_0^t e^{(t-s)A}B\hat{u}(s, h) ds.$$

It follows that for any $T > 0$ there exists $C_T > 0$ such that

$$(5.23) \quad |\hat{x}(t, h)|_H^2 \leq C_T |h|_H^2, \quad 0 \leq t \leq T.$$

Moreover, from (5.21), $\phi(\hat{x}(t, h)) \leq \phi(h)$ and the continuity of $S_\Sigma(t)$ follows. We now prove that $\lim_{t \rightarrow 0} \hat{x}(t, h) = h, \forall h \in \Sigma$. By (5.22) we have $\lim_{t \rightarrow 0} \hat{x}(t, h) = h$ in H . It remains to show that $\hat{x}(t)$ is continuous at $t = 0$ with respect to the seminorm $\phi(h)^{1/2}$. By the linearity of $\hat{x}(\cdot, h)$ and $\hat{u}(\cdot, h)$ in h , we have

$$\phi(\hat{x}(t, h) - h) = \int_0^\infty \{|C\hat{x}(t+s, h) - C\hat{x}(s, h)|^2 + |\hat{u}(t+s, h) - \hat{u}(s, h)|^2\} ds.$$

Since $C\hat{x}(\cdot, h) \in L^2(0, \infty; Y)$ and $\hat{u}(\cdot, h) \in L^2(0, \infty; U)$, we have $\lim_{t \rightarrow 0} \phi(\hat{x}(t, h) - h) = 0$. This proves (iii).

(iv) For any $h \in D(A_\Sigma)$, $\hat{x}(\cdot, h) \in C^1([0, \infty; \Sigma])$ and $\hat{x}'(0, h) = A_\Sigma h$. Denote by $\hat{w}(\cdot) = \hat{u}(\cdot, A_\Sigma h)$ the optimal control corresponding to $A_\Sigma h$. So for all $t > 0$

$$\begin{aligned} \phi \left[\frac{\hat{x}(t, h) - h}{t} - A_\Sigma h \right] &= \int_0^\infty \left\{ \left| \frac{C\hat{x}(t+s, h) - C\hat{x}(s, h)}{t} - C\hat{x}'(s, h) \right|^2 \right. \\ &\quad \left. + \left| \frac{\hat{u}(t+s, h) - \hat{u}(s, h)}{t} - \hat{w}(s) \right|^2 \right\} ds. \end{aligned}$$

As t goes to zero, the first two terms go to zero and necessarily

$$\lim_{t \rightarrow 0} \int_0^\infty \left| \frac{\hat{u}(t+s) - \hat{u}(s)}{t} - \hat{w}(s) \right|^2 ds = 0,$$

which implies $\hat{w} = \hat{u}'$ and $\hat{u} \in H^1(0, \infty; U)$, $\forall h \in D(A_\Sigma)$. By (5.22) we conclude that $h \in D(A)$, and (5.6) follows.

(v) We have shown in (ii) that the map $h \rightarrow \hat{u}(\cdot, h) : \Sigma \rightarrow L^2(0, \infty; U)$ is linear and continuous. In particular,

$$h \rightarrow \hat{u}'(\cdot, h) = \hat{u}(\cdot, A_\Sigma h) : D(A_\Sigma) \rightarrow L^2(0, \infty; U)$$

is also continuous. Hence

$$h \rightarrow \hat{u}(\cdot, h) : D(A_\Sigma) \rightarrow H^1(0, \infty; U)$$

is linear and continuous when $D(A_\Sigma)$ is endowed with the following graph norm topology:

$$\|h\|_{D(A_\Sigma)}^2 = \|h\|_\Sigma^2 + \|A_\Sigma h\|^2.$$

In particular, $\hat{u}(\infty) = 0$, $\hat{u} \in C([0, \infty]; U)$ and the map $h \rightarrow \hat{u}(0, h) : D(A_\Sigma) \rightarrow U$ is linear and continuous. We denote it by K . Equivalently, K is a closed linear unbounded operator from Σ to U with domain

$$D(K) = \{h \in \Sigma : Kh \in U\} \supset D(A_\Sigma).$$

In view of this and identity (5.6)

$$\forall h \in D(A_\Sigma), \quad A_\Sigma h = Ah + B\hat{u}(0, h) = [A + BK]h.$$

Conversely, if $h \in D(A) \cap D(K)$, then

$$A_\Sigma h = Ah + BK h \Rightarrow h \in D(A_\Sigma),$$

and $D(A_\Sigma) = D(A) \cap D(K)$.

(vi) To relate K and the limit of $P(t)$, we go back to formula (4.2) with $h \in \Sigma$, $u = \hat{u}(\cdot, h)$ and $x = \hat{x}(\cdot, h)$:

$$\begin{aligned} (5.24) \quad \langle P(t)h, h \rangle_\Sigma &+ \int_0^t |\hat{u}(s, h) + B^*P(t-s)\hat{x}(s, h)|^2 ds \\ &= \int_0^t \{|C\hat{x}(s, h)|^2 + |\hat{u}(s, h)|^2\} ds. \end{aligned}$$

As t goes to infinity, we obtain

$$\lim_{t \rightarrow \infty} \int_0^t |\hat{u}(s, h) + B^*P(t-s)\hat{x}(s, h)|^2 ds = 0.$$

Setting $P(r) = 0$ for $r \leq 0$, then

$$(5.25) \quad \lim_{t \rightarrow \infty} \int_0^\infty |\hat{u}(s, h) + B^*P(t-s)\hat{x}(s, h)|^2 ds = 0,$$

since

$$\lim_{t \rightarrow \infty} \int_t^\infty |\hat{u}(s, h)|^2 ds = 0.$$

Now repeat the same estimate with $A_\Sigma h$ instead of h and $\hat{x}(\cdot, A_\Sigma h) = \hat{x}'(\cdot, h)$, $\hat{u}(\cdot, A_\Sigma h) = \hat{u}'(\cdot, h)$. Then by the same argument

$$(5.26) \quad \lim_{t \rightarrow \infty} \int_0^\infty |\hat{u}'(s, h) + B^*P(t-s)\hat{x}'(s, h)|^2 ds = 0.$$

Introduce the notation, and use (5.25) and (5.26) as follows:

$$\begin{aligned} v_t(s) &= \hat{u}(s, h) + B^*P(t-s)\hat{x}(s, h), & v_t &\rightarrow 0 \quad \text{in } L^2(0, \infty; U), \\ w_t(s) &= \hat{u}'(s, h) + B^*P(t-s)\hat{x}'(s, h), & w_t &\rightarrow 0 \quad \text{in } L^2(0, \infty; U). \end{aligned}$$

For h in $D(A_\Sigma)$, differentiate (5.24) with respect to t

$$\frac{d}{dt} \langle P(t)h, h \rangle + |\hat{u}(0, h) + B^*P(t)h|^2 + 2 \int_0^t (v_t(s), w_t(s)) ds = |C\hat{x}(t, h)|^2 + |\hat{u}(t, h)|^2.$$

For $t' \geq t$, however,

$$\langle P(t')h, h \rangle - \langle P(t)h, h \rangle \geq 0 \Rightarrow \frac{d}{dt} \langle P(t)h, h \rangle \geq 0,$$

and note that

$$\lim_{t \rightarrow \infty} \int_0^t (v_t(s), w_t(s)) ds = \lim_{t \rightarrow \infty} \int_0^\infty (v_t(s), w_t(s)) ds \rightarrow 0 \quad \text{as } t \rightarrow \infty.$$

Hence

$$\begin{aligned} 0 &\leq \limsup_{t \rightarrow \infty} \frac{d}{dt} \langle P(t)h, h \rangle \leq \limsup_{t \rightarrow \infty} \{|C\hat{x}(t, h)|^2 + |\hat{u}(t, h)|^2\}, \\ 0 &\leq \limsup_{t \rightarrow \infty} |\hat{u}(0, h) + B^*P(t)h|^2 \leq \limsup_{t \rightarrow \infty} \{|C\hat{x}(t, h)|^2 + |\hat{u}(t, h)|^2\}, \end{aligned}$$

and the \liminf are positive. Recall, however, that $C\hat{x}(\cdot, h) \in H^1(0, \infty; Y)$ and $\hat{u}(\cdot, h) \in H^1(0, \infty; U)$, and this implies that $\lim_{t \rightarrow \infty} C\hat{x}(\cdot, h) = 0$ and $\lim_{t \rightarrow \infty} \hat{u}(\cdot, h) = 0$, and the limit of the two terms exists and is equal to 0. So, finally, for all h in $D(A_\Sigma)$ $Kh = \lim_{t \rightarrow \infty} [-B^*P(t)h]$. \square

Remark 5.1. Theorem 5.1 shows that

$$(5.27) \quad V(h) = \phi(h) = V(\hat{x}(0, h)) = \int_0^\infty \{|C\hat{x}(s, h)|^2 + |\hat{u}(s, h)|^2\} ds.$$

Hence $\text{dom } V = \text{dom } \phi = \Sigma$, and Σ coincides with the domain of stabilization of the triple (A, B, C) .

Moreover, by the linearity of $\hat{x}(s, h)$ and $\hat{u}(s, h)$ in h , it follows that

$$(5.28) \quad \psi(h, k) = \int_0^\infty \{(C\hat{x}(s, h), C\hat{x}(s, k)) + (\hat{u}(s, h), \hat{u}(s, k))\} ds, \quad \forall h, k \in \Sigma.$$

6. The algebraic Riccati equation. Recall that we have identified the elements of the dual H' of H with those of H . Our first task is to give a meaning to a solution of the operator algebraic Riccati equation. Let Q be a positive selfadjoint closed linear operator from H to H with a dense domain $D(Q)$. Define

$$(6.1) \quad \Sigma_Q = D(Q^{1/2}) \text{ endowed with its graph norm topology,}$$

$$(6.2) \quad A_Q = A - BB^*Q \text{ on } D(A) \cap D(Q), \text{ and}$$

$$(6.3) \quad \bar{A}_Q = \text{closure of } A_Q \text{ in } \Sigma_Q \text{ (closure of the graph of } A_Q \text{ in } \Sigma_Q \times \Sigma_Q).$$

DEFINITION 6.1. We say that a positive selfadjoint closed linear operator Q with dense domain in H is a solution of the operator algebraic Riccati equation if

- (i) \bar{A}_Q is the infinitesimal generator of a strongly continuous semigroup $\{S_Q(t)\}$ of class C_0 on Σ_Q , and
- (ii) Q verifies the following equation:

$$(6.4) \quad (Qh, Ak) + (Qk, Ah) - (B^*Qh, B^*Qk)_U + (Ch, Ck)_Y = 0, \quad \forall h, k \in D(A) \cap D(Q).$$

DEFINITION 6.2. We say that the triplet (A, B, C) is *approximately stabilizable* (respectively, *stabilizable*) if $\bar{\Sigma} = H$ (respectively, $\Sigma = H$).

Remark 6.1. Note that our definition of approximate stabilizability does not assume the existence of a bounded linear feedback operator. In the literature on the control of infinite dimensional systems, many papers use a definition of stabilizability that assumes the existence of a bounded feedback operator (cf., for instance, Jacobson and Nett [8]). As we will see in Example 6.1, there are simple control systems for which there exists only an unbounded feedback operator, which makes the closed loop system stable for all initial conditions in the state space H . So for infinite dimensional control systems a hypothesis using the existence of a bounded feedback really contains two hypotheses in one. To clarify this question we would have to systematically go over this literature. However, this is not the objective of this paper.

Proposition 6.1. (i) If the triplet (A, B, C) is approximately stabilizable, then the operator P_∞ on H defined by (4.15) is a positive selfadjoint closed linear solution of the operator algebraic Riccati equation (6.4). Moreover, for any other positive selfadjoint closed linear solution Q to (6.4), P_∞ is the minimum solution, that is,

$$(6.5) \quad D(Q^{1/2}) \subset D(P_\infty^{1/2}), \quad \text{and} \quad \forall h \in D(Q), (Qh, h) \geq (P_\infty h, h).$$

(ii) The operator algebraic Riccati equation (6.4) has a positive selfadjoint solution in the sense of Definition 6.1 if and only if the triplet (A, B, C) is approximately stabilizable.

Proof. (i) Recall that from (5.28) for all h and k in $D(A_\Sigma)$

$$\begin{aligned} \langle P_\infty A_\Sigma h, k \rangle &= \int_0^\infty \{(C\hat{x}'(s, h), C\hat{x}(s, k))_Y + (\hat{u}'(s, h), \hat{u}(s, k))_U\} ds, \\ \langle P_\infty h, A_\Sigma k \rangle &= \int_0^\infty \{(C\hat{x}(s, h), C\hat{x}'(s, k))_Y + (\hat{u}(s, h), \hat{u}'(s, k))_U\} ds. \end{aligned}$$

Now $C\hat{x}(\cdot, h)$ and $C\hat{x}(\cdot, k)$ belong to $H^1(0, \infty; Y)$; $\hat{u}(\cdot, h)$ and $\hat{u}(\cdot, k)$ belong to $H^1(0, \infty; Y)$, and their limits as t goes to infinity are 0. Therefore

$$\begin{aligned} \langle P_\infty A_\Sigma h, k \rangle + \langle P_\infty h, A_\Sigma k \rangle &= \int_0^\infty \frac{d}{dt} \{ (C\hat{x}(s, h), C\hat{x}(s, k))_Y + (\hat{u}(s, h), \hat{u}(s, k))_U \} ds \\ &= -(Ch, Ck)_Y - (\hat{u}(0, h), \hat{u}(0, k))_U. \end{aligned}$$

In view of expression (5.10) to (5.12) in Theorem 5.1 we readily obtain (6.4) by specializing to h and k in $D(Q) \cap D(A)$.

Let Q be another positive selfadjoint solution of the operator algebraic Riccati equation (6.4). Then we can rearrange the terms in the following way:

$$\begin{aligned} ([A - BB^*Q]h, Qk) + (Qh, [A - BB^*Q]k) + (B^*Qh, B^*Qk)_U + (Ch, Ck)_Y &= 0, \\ \forall h, k \in D(Q) \cap D(A). \end{aligned}$$

By hypothesis

$$(6.6) \quad (\bar{A}_Q h, Qk) + (Qh, \bar{A}_Q k) + (B^*Qh, B^*Qk)_U + (Ch, Ck)_Y = 0$$

and

$$(B^*Qh, B^*Qk)_U = -[(Ch, Ck)_Y - (Q^{1/2}\bar{A}_Q h, Q^{1/2}k) - (Q^{1/2}h, Q^{1/2}\bar{A}_Q k)].$$

However, $D(Q) \cap D(A) \subset D(\bar{A}_Q)$ and, by linearity and density, the above equation extends to all h and k in $D(\bar{A}_Q)$. In particular, the operator $K_Q = -B^*Q$ has a continuous linear extension $\bar{K}_Q: D(\bar{K}_Q) \subset H \rightarrow U$ such that $D(\bar{K}_Q) \supset D(\bar{A}_Q)$.

For all h in $D(\bar{A}_Q)$,

$$2(\bar{A}_Q S_Q(t)h, Qh) + |B^*Q S_Q(t)h|_U^2 + |C S_Q(t)h|_Y^2 = 0$$

and for all $t \geq 0$

$$|Q^{1/2} S_Q(t)h|^2 + \int_0^t \{ |B^*Q S_Q(s)h|_U^2 + |C S_Q(s)h|_Y^2 \} ds = |Q^{1/2}h|^2.$$

Therefore

$$(6.7) \quad \forall t \geq 0, \quad \forall h \in D(\bar{A}_Q), \quad \int_0^t \{ |u_Q(s)|_U^2 + |C x_Q(s)|_Y^2 \} ds \leq |Q^{1/2}h|^2,$$

where

$$u_Q(s) = -B^*Q x_Q(s) \quad \text{and} \quad x_Q(s) = S_Q(s)h, \quad s \geq 0.$$

Using the monotone increasing property of the integral, inequality (6.7) holds with $t = \infty$ and extends to all h in $D(Q^{1/2})$. Recall that for all h in $\Sigma = D(P_\infty^{1/2})$

$$\int_0^\infty \{ |C\hat{x}(s, h)|^2 + |\hat{u}(s, h)|^2 \} ds = |P_\infty^{1/2}h|^2$$

for the control, and state

$$\hat{u}(s, h) = -B^*Q\hat{x}(s, h) \quad \text{and} \quad \hat{x}(s, h) = S_\Sigma(s)h, \quad s \geq 0.$$

Hence, by minimality of the optimal control $\hat{u}(\cdot, h)$,

$$|P_\infty^{1/2}h|^2 = J(\hat{u}(\cdot, h), h) \leq J(u_Q, h) \leq |Q^{1/2}h|^2$$

and, necessarily, $D(Q^{1/2}) \subset D(P_\infty^{1/2}) = \Sigma$.

(ii) From part (i) we have already established that (6.4) has a positive selfadjoint solution if (A, B, C) is stabilized. Conversely, if Q is a positive selfadjoint solution to the operator algebraic Riccati equation (6.4), then we can repeat the step in part (i) and obtain (6.7), which says that the dense subset $D(\bar{A}_Q)$ of initial conditions is (A, B, C) stabilizable. In particular, $D(\bar{A}_Q) \subset \Sigma$ and $\bar{\Sigma} = H$. \square

Example 6.1. Recall the example in § 3. We have seen that

$$(6.8) \quad H = D(A) = \ell^2, \quad \Sigma = \left\{ h \in \ell^2: \sum_{k=1}^{\infty} (k+1)h_k^2 < \infty \right\}, \quad \bar{\Sigma} = H,$$

$$(6.9) \quad D(P_{\infty}) = \left\{ h \in \ell^2: \sum_{k=1}^{\infty} (k+1)^2 h_k^2 < \infty \right\}, \quad P_{\infty} e_k = (k+1)e_k, \quad k \in \mathbb{N}.$$

Moreover, $D(A_{\Sigma}) = \Sigma$, and K is the closed operator in H

$$(6.10) \quad D(K) = \Sigma, \quad K e_k = \sqrt{2k+1} e_k, \quad k \in \mathbb{N}.$$

The space Σ is the set of all initial conditions that can be stabilized with a finite energy. However, for all h in H

$$(6.11) \quad \int_0^{\infty} |x(s)|_H^2 ds < \infty,$$

and for all h in Σ

$$(6.12) \quad \int_0^{\infty} |x(s)|_{\Sigma}^2 ds = \int_0^{\infty} \{ |x(s)|_H^2 + \langle P_{\infty} x(s), x(s) \rangle_{\Sigma} \} ds \leq c|h|^2.$$

We remark that, in general, the closed loop system is not exponentially stable, as the following example shows.

Example 6.2. Let $H = D(A) = \ell^2$,

$$(6.13) \quad A e_k = 0, \quad B e_k = \frac{1}{k} e_k, \quad C e_k = e_k.$$

Then

$$(6.14) \quad P_{\infty} e_k = k e_k, \quad k \in \mathbb{N},$$

$$(6.15) \quad \Sigma = \left\{ h \in \ell^2: \sum_{k=1}^{\infty} k h_k^2 < \infty \right\}, \quad \bar{\Sigma} = H,$$

$$(6.16) \quad F e_k = (A - B B^* P_{\infty}) e_k = -\frac{1}{k} e_k.$$

Thus F is stable but not exponentially stable both in H and in Σ .

PROPOSITION 6.2. *If the triplet (A, B, C) is approximatively stabilizable and the pair (A^*, C^*) is stabilizable, then*

$$(6.17) \quad \int_0^{\infty} |\hat{x}(t, h)|_H^2 dt < \infty, \quad \text{for all } h \in \Sigma$$

and the triplet (A, B, I) is approximatively stabilizable.

Proof. If (A^*, C^*) is stabilizable, then there exists a minimum positive bounded solution to the Riccati equation

$$(6.18) \quad A Q + Q A^* - Q C^* C Q + I = 0$$

and the closed loop system

$$(6.19) \quad y'(t) = [A^* - C^*CQ]y(t), \quad y(0) = k$$

is L^2 -stable. Denote by $T(t)$ the semigroup associated with the above system. For all h in Σ consider the optimal state $\hat{x}(\cdot, h)$ and control $\hat{u}(\cdot, h)$, then

$$(6.20) \quad \hat{x}'(t, h) = [A^* - C^*CQ]^* \hat{x}(t, h) + QC^*C\hat{x}(t, h) + B\hat{u}(t, h), \quad \hat{x}(0, h) = h$$

and

$$(6.21) \quad \hat{x}(t, h) = T^*(t)h + \int_0^t T^*(t-s)[QC^*C\hat{x}(s, h) + B\hat{u}(s, h)] ds.$$

It follows that

$$\begin{aligned} \|\hat{x}(t; h)\|_{L^2(0, \infty; H)} &\leq \|T^*(\cdot)h\|_{L^2(0, \infty; H)} + \|T^*(\cdot)h\|_{L^2(0, \infty; H)} \|QC^*C\hat{x}(\cdot, h) \\ &\quad + B\hat{u}(\cdot, h)\|_{L^2(0, \infty; H)}. \end{aligned}$$

The right-hand side is finite since T^* is exponentially decreasing, QC^* and B are bounded, and $C\hat{x}(\cdot, h)$ and $\hat{u}(\cdot, h)$ are $L^2(0, \infty; H)$ functions. \square

Remark 6.2. To show the L^2 -stability with respect to the Σ norm, we would have to prove that

$$(6.22) \quad \int_0^\infty \langle P_\infty \hat{x}(t; h), \hat{x}(t, h) \rangle_\Sigma dt = \int_0^\infty \int_t^\infty [|C\hat{x}(s, h)|^2 + |\hat{u}(s, h)|^2] ds dt < \infty.$$

7. A condition for approximative stabilizability. In this section we examine the connection between the (A, B, I) approximative stabilizability and the Hautus condition. We present a set of conditions (Hypothesis 7.1) under which the equivalence is verified (Proposition 7.1). We complete this section with an application of Hypothesis 7.1 to the nerve axon system (Example 7.1).

Hypothesis 7.1. Let A be the infinitesimal generator of an analytic semigroup on H . Denote by $\sigma(A)$ the spectrum of A , and by $\rho(A)$ the resolvent set of A . Assume that the following properties are verified:

- (i) $\sigma(A)$ consists of a convergent sequence $\{\lambda_i\}$ of semisimple² eigenvalues plus the limit point $\lambda_\infty = \lim_{i \rightarrow \infty} \lambda_i$;
- (ii) $\sigma(A) = \sigma^-(A) \cup \sigma^+(A)$, where $\sigma^-(A) = \{\lambda: \operatorname{Re} \lambda < 0\}$ and $\sigma^+(A) = \{\lambda: \operatorname{Re} \lambda > 0\}$. We set $P_+ = 1/(2\pi i) \int_\gamma (\lambda - A)^{-1} d\lambda$, where γ is a suitable curve around $\sigma^+(A)$ and define $P = I - P_+$;
- (iii) Setting $P_i = 1/(2\pi i) \int_{C(\lambda_i, \varepsilon_i)} (\lambda - A)^{-1} d\lambda$, where $C(\lambda_i, \varepsilon_i)$ is a circle in $\rho(A)$, we have $e^{tA}P_+x = \sum_{i=1}^\infty e^{t\lambda_i}P_i x$.

PROPOSITION 7.1. Assume that Hypothesis 7.1 is verified and that $B \in \mathcal{L}(U; H)$. Then the following statements are equivalent:

- (i) The triple (A, B, I) is approximatively stabilizable,
- (ii) $\operatorname{Ker}(B^*) \cap \operatorname{Ker}(A^* - \lambda_i I) = \{0\}$ for all $\lambda_i \in \sigma^+(A)$.

Proof. (i) \Rightarrow (ii). Assume, by contradiction, that (A, B, I) is approximatively stabilizable and that there exists $\lambda \in \sigma(A)$ and h in H , $|h| = 1$, such that $A^*h = \bar{\lambda}h$, $B^*h = 0$. By (A, B, I) approximative stabilizability for any k in Σ there exists a control u in $L^2(0, \infty; U)$ such that the corresponding solution x of (2.1) belongs to $L^2(0, \infty; U)$. Define the function $g(t) = (h, x(t))$. Then g is the solution of the equation

$$g'(t) = \lambda g(t), \quad t \geq 0, \quad g(0) = (h, k) \Rightarrow g(t) = (h, k) e^{\lambda t}, \quad t \geq 0.$$

² An eigenvalue is said to be semisimple if it is an isolated point of the spectrum and a simple pole of the resolvent operator (cf. T Kato [4, p. 41]).

Hence

$$|(h, k)| \int_0^\infty e^{\operatorname{Re} \lambda t} dt = \|g\|_{L^2(0, \infty)} \leq \|h\|_{L^2(0, \infty; U)} < \infty.$$

However,

$$(7.1) \quad \int_0^\infty e^{\operatorname{Re} \lambda t} dt = \infty \Rightarrow \forall k \in \Sigma, \quad (h, k) = 0,$$

and by density of Σ in H , $h = 0$, which is in contradiction with our hypothesis.

(ii) \Rightarrow (i). Let $h \in H$ and $u \in L^2(0, \infty; U)$. We can write the solution of problem (2.1) as

$$(7.2) \quad \begin{aligned} x(t) = & e^{tA} Ph + \int_0^t e^{(t-s)A} P_- Bu(s) ds - \int_t^\infty e^{(t-s)A} P_+ Bu(s) ds \\ & + e^{tA} \left\{ P_+ h + \int_0^\infty e^{-sA} P_+ Bu(s) ds \right\}. \end{aligned}$$

Thus the control u is admissible if and only if $P_+ h + \int_0^\infty e^{-sA} P_+ Bu(s) ds = 0$. Consider now the mapping

$$(7.3) \quad u \rightarrow \gamma(u) = \int_0^\infty e^{-sA} P_+ Bu(s) ds : L^2(0, \infty; U) \rightarrow P_+ H = H_+$$

and its adjoint

$$(7.4) \quad h \rightarrow (\gamma^* h)(s) = B^* e^{-sA^*} h : H_+^* \rightarrow L^2(0, \infty; U).$$

Clearly the triple (A, B, I) is approximatively stabilizable if and only if $\operatorname{Ker}(\gamma^*) = \{0\}$. Now assume that (ii) holds and, by contradiction, that $\operatorname{Ker}(\gamma^*) \neq \{0\}$. In view of Hypothesis 7.1 (iii) for any $h \in \operatorname{Ker}(\gamma^*)$ we have

$$(7.5) \quad B^* e^{-sA^*} h = B^* \sum_{i=1}^\infty e^{-t\lambda_i} P_i^* h = 0,$$

which implies $P_i^* h \in \operatorname{Ker}(B^*)$. Since $P_i^* h \in \operatorname{Ker}(A^* - \lambda_i I)$ (because λ_i is semisimple) we have found a contradiction with (ii). \square

Example 7.1 (The nerve axon system). Let Ω be an open bounded set in \mathbf{R} and consider the system (introduced in [3])

$$(7.6) \quad \begin{aligned} \frac{\partial x_1}{\partial t}(t, \xi) &= \alpha \Delta x_1(t, \xi) + b_{11} x_1(t, \xi) + b_{12} x_2(t, \xi) + \sum_{j=1}^J f_j(t) \phi_j(t, \xi), \\ t > 0, \quad \xi \in \Omega, \\ \frac{\partial x_2}{\partial t}(t, \xi) &= b_{21} x_1(t, \xi) + b_{22} x_2(t, \xi) + \sum_{j=1}^J g_j(t) \psi_j(t, \xi), \quad t > 0, \quad \xi \in \Omega, \\ x_1(0, x) &= h_1(x), \quad \xi \in \Omega, \\ x_2(0, x) &= h_2(x), \quad \xi \in \Omega, \\ x_1(t, \xi) &= 0, \quad x_2(t, \xi) = 0, \quad t > 0, \quad \xi \in \partial\Omega, \end{aligned}$$

where we assume that $\alpha, b_{ij} : \mathbf{R} \rightarrow \mathbf{R}$ are given real numbers, with $\alpha > 0$, $b_{12} b_{21} \neq 0$, and $\phi_1, \dots, \phi_J, \psi_1, \dots, \psi_J \in C(\bar{\Omega})$ are linearly independent functions.

Choose $H = L^2(\Omega) \times L^2(\Omega)$, $U = \mathbf{R}^J \times \mathbf{R}^J$. Setting

$$(7.7) \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad h = \begin{bmatrix} h_1 \\ h_2 \end{bmatrix}, \quad u = \begin{bmatrix} (f_1, \dots, f_J) \\ (g_1, \dots, g_J) \end{bmatrix}, \quad Bu = \begin{bmatrix} \sum_{j=1}^J f_j \phi_j(t, \cdot) \\ \sum_{j=1}^J g_j \psi_j(t, \cdot) \end{bmatrix}$$

and

$$(7.8) \quad b = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix},$$

we can write system (7.6) in the abstract form (2.1). The spectrum $\sigma(A)$ of A consists in two sequences of semisimple eigenvalues $\{\lambda_{\pm}(k)\}_{k \in \mathbf{N}}$ and the accumulation point

$$(7.10) \quad \lambda_{\infty} = b_{22}.$$

The eigenvalues $\lambda_{\pm}(k)$ are defined by

$$(7.11) \quad \lambda_{\pm}(k) = \frac{1}{2}[-\alpha\mu_k + \text{Tr}(b) \pm \sqrt{[-\alpha\mu_k + \text{Tr}(b)]^2 + 4[\alpha\mu_k b_{22} - \det(b)]}],$$

where the μ_k 's are the eigenvalues of the Laplacian with Dirichlet boundary conditions. Now it is easy to check Hypothesis 7.1, so that we can apply Proposition 7.1.

REFERENCES

- [1] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Springer-Verlag, Berlin, Heidelberg, New York, 1978.
- [2] G. DA PRATO AND M. C. DELFOUR, *Stabilization and unbounded solutions of the Riccati equation*, in Proc 27th IEEE Conference on Decision and Control, IEEE Publ., New York, 1988, pp. 352-357.
- [3] J. EVANS, *Nerve axon equations: I linear approximations; II stability at rest; III stability of the nerve impulse*, Indiana University Math. J., 21 (1972), pp. 877-885; 22 (1972), pp. 75-90; 22 (1972), pp. 577-593.
- [4] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
- [5] J. L. LIONS AND E. MAGENES, *Problèmes aux limites non homogènes et applications*, Vols. 1, 2, 3, Dunod, Paris, 1968, 1969, 1970.
- [6] J. L. LIONS AND J. PEETRE, *Sur une classe d'espaces d'interpolation*, Pub. Math. de l'I.H.E.S., 19 (1964), pp. 5-68.
- [7] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [8] C. A. JACOBSON AND C. N. NETT, *Linear state-space systems in infinite-dimensional space: the role and characterization of joint stabilizability/detectability*, IEEE Trans. Automat. Control, AC 33 (1988), pp. 541-549.

LOCAL CONVERGENCES AND OPTIMAL SHAPE DESIGN*

WENBIN LIU† AND J. E. RUBIO‡

Abstract. Several new concepts dealing with the convergence of convex sets and functionals in various spaces are introduced. Some compactness and lower-semicontinuity results with respect to the convergences are established. Then three general existence results of optimal shapes for variational inequalities are obtained.

Key words. optimal shape design, local convergence, existence theorems, variational convergence, variational inequality

AMS(MOS) subject classification. 49A29

1. Introduction. In [5]–[8] and [11]–[13], some existence results of optimal shape design for partial differential equations and variational inequalities have been established. In that work, the original problem is transformed to establish existence results. However, it is now realized that we can study domain optimization problems directly by introducing some concepts dealing with the convergences in various spaces, and thus establish more general results.

The plan of our paper is as follows. In § 2 we introduce some general results that will be used later, and we treat the local convergences of functionals in various spaces. Then some compactness and lower-semicontinuity results with respect to the convergences are established. In § 3 we introduce and study some new Mosco convergences of closed convex sets in various spaces. Finally, in § 4, as examples of some applications of the new concepts, we establish some general existence results of optimal shapes for variational inequalities, which can cover some results in [5]–[8] and [11]–[13].

2. Notation and preliminaries. In this section, we introduce and prove some results that will be used later.

Let A and B be two subsets of R^m and define δ as follows (see [13]):

$$\delta(A, B) = \max \{ \rho(A, B), \rho(B, A) \}, \quad \text{where } \rho(A, B) = \sup_{x \in A} \inf_{y \in B} \|x - y\|_{R^m}.$$

Then it is well known that δ is a metric on the closed subsets of R^m , the Hausdorff metric. A very useful property is presented in the following theorem.

THEOREM 2.1 ([13]). *Let $\{A_n\}$ be a sequence of closed sets of R^m such that $A_n \subset C$, a bounded closed set of R^m . Then there are a subsequence of $\{A_n\}$, still denoted as $\{A_n\}$, and a closed set $A \subset C$ such that $\delta(A_n, A) \rightarrow 0$.*

By using the metric δ , we can also study the convergence of open sets [13].

DEFINITION 2.1. Let $\{\Omega_n\}$ ($n = 0, 1, \dots$) be a sequence of open sets contained in an open set $C \subset R^m$. We say that $\Omega_n \xrightarrow{C} \Omega_0$ if $\delta(\bar{\Omega}_n, \bar{\Omega}_0) \rightarrow 0$, where $\bar{\Omega}^c \equiv \bar{C} \setminus \Omega$ and \bar{C} is the closure of C in R^m .

It follows from Theorem 2.1 that if $\{\Omega_n\}$ ($n = 1, 2, \dots$), a sequence of open sets in R^m , is contained in C bounded open set of R^m , then there are subsequence of $\{\Omega_n\}$, still denoted as $\{\Omega_n\}$, and a closed F in \bar{C} such that $\delta(\bar{C} \setminus \Omega_n, F) \rightarrow 0$. It is easy to see that $\bar{C} \setminus F$ is open because $\bar{C} \setminus F \subset C$ (note that $\bar{C} \setminus C \subset F$ from Lemma 3.1 [3]) and so $R^m \setminus (\bar{C} \setminus F) = F \cup (R^m \setminus \bar{C})$ is closed. Therefore, there is an open set $\Omega_0 = \bar{C} \setminus F \subset C$ such that $\Omega_n \xrightarrow{C} \Omega_0$.

* Received by the editors October 23, 1989; accepted for publication (in revised form) November 30, 1990.

† School of Mathematics, University of Leeds, Leeds LS2 9JT, England.

LEMMA 2.1. Let $\{\Omega_n\}$ ($n=0, 1, \dots$) be a sequence of open sets of R^m , contained in a bounded set of R^m . If $\{\Omega_n\}$ is contained in C , an open set of R^m and $\Omega_n \xrightarrow{C} \Omega_0$, then for any open set $G \Subset \Omega_0$ (that is, \bar{G} is compact and $\bar{G} \subset \Omega_0$), there is $N(G) > 0$ such that $G \subset \Omega_n$ ($n \geq N(G)$) (see the proof of Proposition 3.1 in [13]).

Remark 2.1. We can give another proof for Lemma 2.1 by the fact that $\delta(C \setminus \Omega_n, C \setminus \Omega_0) \geq \rho(x, C \setminus \Omega_0)$ if $x \in G$ and $x \notin \Omega_n$. We also note that C is not necessarily bounded in the above lemma. It follows from this lemma that for any open set $G \Subset \Omega_0$, we find that $N(G) > 0$ such that $G \Subset \Omega_n$ for $n \geq N(G)$ because we find that another open set $G_0 \Subset \Omega_0$ such that $G \Subset G_0$.

We say that the open set sequence $\{\Omega_n\}$ ($n=0, 1, 2, \dots$) has the property G if, for any open set $G \Subset \Omega_0$, there is $N(G) > 0$ such that $G \subset \Omega_n$ ($n \geq N(G)$). Thus, if $\Omega_n \xrightarrow{C} \Omega_0$ ($n=1, 2, \dots$), then $\{\Omega_n\}$ ($n=0, 1, \dots$) has the property G .

It is clear that such a convergence is too weak, as it cannot even guarantee that the Lebesgue measure of Ω_n tends to that of Ω_0 when $\Omega_n \xrightarrow{C} \Omega_0$, and, furthermore, we cannot get much information about $\partial\Omega_0$ from what we know about $\{\partial\Omega_n\}$. So we introduce a stronger convergence.

THEOREM 2.2 ([5], [13]). Let $O_\varepsilon = \{\Omega \text{ open, } \Omega \subset C \text{ a fixed bounded open of } R^m, \text{ and } \Omega \text{ has the } \varepsilon\text{-cone property}\}$ (see [5] and [13]) for some $\varepsilon > 0$. Then O_ε is a compact metric space where the following distance is defined:

$$d_C(A, B) = \delta(A^\varepsilon, B^\varepsilon) + \left(\int_{R^m} |\chi_A - \chi_B| dx \right),$$

where χ_A is the characteristics function of A .

Note that the compactness and the property in Lemma 2.1 for O_ε can be established without introduction of δ (see [5] and [6]).

THEOREM 2.3 ([5]). If $\Omega \in O_\varepsilon$, then it has a Lipschitz boundary and there is an extension operator j_Ω from $H^1(\Omega)$ to $H^1(R^m)$ such that

$$\|j_\Omega u\|_{H^1(R^m)} \leq c \|u\|_{H^1(\Omega)}, \quad \text{for any } u \in H^1(\Omega),$$

where the constant c is independent of Ω .

Remark 2.2. Let C be a fixed open set in R^m with a $W^{r,\infty}$ boundary (see [11]). We can introduce (see [11] and [12]), for $k \geq 1$,

$$O^{k,\infty} \equiv \{T(C); T \in \mathcal{F}^{k,\infty}\},$$

where $\mathcal{F}^{k,\infty} \equiv \{T; T \text{ is a bijective from } R^m \text{ to } R^m \text{ and } I - T, I - T^{-1} \in [W^{k,\infty}(R^m)]^m\}$. It can be shown that for any $k \geq 1$ there is a topology τ on $O^{k,\infty}$, which is finer than the topology given by the distance d_{R^m} in Theorem 2.2, such that any bounded closed sets in $O^{k,\infty}$ are compact in $O^{k-1,\infty}$ (see [11] and [12]). There are some other classes of open sets with topologies finer than that introduced by the distance d_C such that the topology spaces are compact. These classes do not belong to O_ε or $O^{k,\infty}$.

Finally, we introduce some concepts of local convergences that are similar to those in [14] and prove some results on compactness and lower-semicontinuity with respect to the convergences.

DEFINITION 2.2. Assume that $\{\Omega_n\}$ is a bounded set sequence in R^m with the property G and that $u_n \in H^1(\Omega_n)$ for $n=0, 1, 2, \dots$. If the sequence $\{\|u_n\|_{H^1(\Omega_n)}\}$ is bounded and for any open set $G \Subset \Omega_0$, there is $N(G) > 0$ such that $u_n \in H^1(G)$ ($n \geq N$) and $\|u_n - u_0\|_{L^2(G)} \rightarrow 0$ ($\|u_n - u_0\|_{H^1(G)} \rightarrow 0$), then we say that $\{u_n\}$ locally weakly (strongly) converges to u_0 in $H^1(\Omega_0)$, or

$$u_n \xrightarrow{L} u_0 \text{ weakly (strongly) in } H^1(\Omega_0).$$

THEOREM 2.4. *Let $\{\Omega_n\}$ be a sequence of bounded open sets of R^m with the property G and $u_n \in H^1(\Omega_n)$ for $n = 0, 1, \dots$. If for any fixed open set $G \Subset \Omega_0$, the functional $u \rightarrow \int_G h(x, u, \nabla u) dx$ is weakly lower-semicontinuous (l.s.c.) in $H^1(G)$, where h is a nonnegative function measurable on $R^m \times R \times R^m$, then*

$$\int_{\Omega_0} h(x, u_0, \nabla u_0) dx \leq \liminf_{n \rightarrow \infty} \int_{\Omega_n} h(x, u_n, \nabla u_n) dx,$$

if $u_n \xrightarrow{L} u_0$ weakly in $H^1(\Omega_0)$.

Proof. Let $\{G_j\}$ be a sequence of bounded open sets such that $G_j \Subset \Omega_0$, $G_j \subset G_{j+1}$, and $\bigcup_1^\infty G_j = \Omega_0$. Then we see that $\chi_{G_j} \rightarrow \chi_{\Omega_0}$ almost everywhere in R^m . Let $u \xrightarrow{L} u_0$ weakly in $H^1(\Omega_0)$. First, for any fixed G_j , we prove that

$$(2.1) \quad \int_{G_j} h(x, u_0, \nabla u_0) dx \leq \liminf_{n \rightarrow \infty} \int_{G_j} h(x, u_n, \nabla u_n) dx.$$

To this end, let $\{u_{n(k)}\}$, a subsequence of $\{u_n\}$, satisfy

$$\lim_{k \rightarrow \infty} \int_{G_j} h(x, u_{n(k)}, \nabla u_{n(k)}) dx = \liminf_{n \rightarrow \infty} \int_{G_j} h(x, u_n, \nabla u_n) dx.$$

Then we suppose that $u_{n(k)} \rightarrow u_0$ weakly in $H^1(G_j)$ as $\{\|u_n\|_{H^1(\Omega_n)}\}$ is bounded and $u_n \rightarrow u_0$ in $L^2(G_j)$ (note that if $G_j \Subset G$, a bounded open set of R^m , and $u_n \rightarrow u_0$ weakly in $H^1(G)$, then $u_n \rightarrow u_0$ strongly in $L^2(G_j)$ without any assumptions on ∂G_j). So we get (2.1) from the l.s.c. assumption. Next, from Fatou's lemma, we have

$$\begin{aligned} \int_{\Omega_0} h(x, u_0, \nabla u_0) dx &= \int_{\Omega_0} \lim_{j \rightarrow \infty} \chi_{G_j} h(x, u_0, \nabla u_0) dx \\ &\leq \liminf_{j \rightarrow \infty} \int_{G_j} h(x, u_0, \nabla u_0) dx \\ &\leq \liminf_{j \rightarrow \infty} \liminf_{n \rightarrow \infty} \int_{G_j} h(x, u_n, \nabla u_n) dx \\ &\leq \liminf_{j \rightarrow \infty} \liminf_{n \rightarrow \infty} \int_{\Omega_n} h(x, u_n, \nabla u_n) dx \\ &= \liminf_{n \rightarrow \infty} \int_{\Omega_n} h(x, u_n, \nabla u_n) dx, \end{aligned}$$

because for fixed G_j , there is $N > 0$ such that $G_j \subset \Omega_n$ ($n \geq N$) and h is non-negative. \square

COROLLARY 2.1. *Let $\{\Omega_n\}$ ($n = 0, 1, \dots$) be a sequence of bounded open sets in R^m with the property G . Let the nonnegative function h be measurable on $R^m \times R \times R^m$, $h(x, \cdot, \cdot)$ be continuous on $R \times R^m$ and $h(x, t, \cdot)$ be convex. Then*

$$\int_{\Omega_0} h(x, u_0, \nabla u_0) dx \leq \liminf_{n \rightarrow \infty} \int_{\Omega_n} h(x, u_n, \nabla u_n) dx,$$

if $u_n \xrightarrow{L} u_0$ weakly in $H^1(\Omega_0)$.

Proof. It is well known that under the assumption of this corollary, the functional $u \rightarrow \int_G h(x, u, \nabla u) dx$ is weakly l.s.c. in $H^1(G)$ for any fixed bounded open set G in R^m . \square

From Theorem 2.4 we also prove some l.s.c. results in [6] and [14] without the continuity assumption in x , which is restrictive in boundary problems. The following result establishes the weak compactness of the local convergence in H^1 .

THEOREM 2.5. *Let Ω_n be an open set contained in a bounded open set C of R^m for $n=0, 1, \dots$. If $\Omega_n \xrightarrow{c} \Omega_0$, then for any sequence $\{u_n\}$ so that $u_n \in H^1(\Omega_n)$ and $\|u_n\|_{H^1(\Omega_n)} \leq d$, there is a subsequence of $\{u_n\}$, still denoted as $\{u_n\}$, and $u_0 \in H^1(\Omega_0)$, such that $u_n \xrightarrow{L} u_0$ weakly in $H^1(\Omega_0)$.*

Proof. Let G_j ($j=1, 2, \dots$) be the open set such that $G_j \Subset \Omega_0$ for $j=1, 2, \dots$, $G_j \subset G_{j+1}$, and $\bigcup G_j = \Omega_0$. For G_1 there is $\{u_{n_k}^1\}$, a subsequence of $\{u_n\}$, and u_0^1 such that $u_{n_k}^1 \rightarrow u_0^1$ strongly in $L^2(G_1)$, $u_{n_k}^1 \in H^1(\Omega_{n_k}^1)$, $u_0^1 \in H^1(G_1)$, and $\|u_0^1\|_{H^1(G_1)} \leq d$ as $G_1 \Subset \Omega_n$ ($n \geq N(G_1)$) from Lemma 2.1 and Remark 2.1. For G_2 , we also have such $\{u_{n_k}^2\}$, a subsequence of $\{u_{n_k}^1\}$, and u_0^2 such that $u_{n_k}^2 \rightarrow u_0^2$ in $L^2(G_2)$ strongly, $\|u_0^2\|_{H^1(G_2)} \leq d$, and $u_0^2|_{G_1} = u_0^1$, and so on. Let $u_0(x) = u_0^j(x)$ where $x \in G_j$. Then u_0 is well defined and in $L^2(\Omega_0)$, since we have $\|u_0\|_{L^2(G_j)} \leq d$ for any $j=1, 2, \dots$. On the other hand, for any $\varphi \in \mathcal{D}(\Omega_0)$ there is j_0 such that $\varphi \in \mathcal{D}(G_{j_0})$, and thus $\langle \nabla u_0, \varphi \rangle = -\langle \nabla u_j, \varphi \rangle$; that is, $\nabla u_0 = \nabla u_j$ when $x \in G_j$ so that $u_0 \in H^1(\Omega_0)$. Thus put $n_k \equiv n_k^k$ and $u_{n_k} \equiv u_{n_k}^k$ for $k=1, 2, \dots$, and we obtain $u_{n_k} \in H^1(\Omega_{n_k})$ such that $u_{n_k} \in H^1(G_j)$ ($n_k \geq N(G_j)$), $u_{n_k} \rightarrow u_0$ in $L^2(G_j)$ and $\|u_0\|_{H^1(G_j)} \leq d$ for any fixed j . Then for any $G \Subset \Omega_0$ such that $\bar{G} \subset \Omega_0 = \bigcup G_j$, $G \subset G_j$ for a fixed j since $\{G_j\}$ is an open cover of \bar{G} , which is a compact set. Thus, we infer that $u_{n_k} \rightarrow u_0$ strongly in $L^2(G)$. \square

We now turn to the evolution case. Let T be a positive number and the spaces $L^2(0, T; H^1(\Omega))$ be the same as in [9]. For fixed $\theta \geq 0$, let $G^\theta(0, T; H^1(\Omega)) = \{u: u \in L^2(0, T; H^1(\Omega)), t^\theta u' \in L^2(0, T; L^2(\Omega)), \theta \geq 0\}$ with the norm

$$\|u\|_{G^\theta(0, T; H^1(\Omega))} = \left[\int_0^T \|u\|_{H^1(\Omega)}^2 dt + \int_0^T \|t^\theta u'\|_{L^2(\Omega)}^2 dt \right]^{1/2};$$

and let $W(0, T) = \{u \in L^2(0, T; H^1(\Omega)); u' \in L^2(0, T; [H^1(\Omega)]^*)\}$ [9], where $u' = du/dt$ is its vector-valued generalized derivative in [9].

DEFINITION 2.3. Let $\{\Omega_n\}$ ($n=0, 1, \dots$) be a sequence of bounded open sets in R^m with the property G . Let $u_n \in L^2(\kappa, T; H^1(\Omega_n))$ ($\kappa \geq 0$) for $n=0, 1, \dots$. We say that $\{u_n\}$ *locally strongly (weakly) converges* to u_0 in the space $L^2(\kappa, T; H^1(\Omega_0))$, or

$$u_n \xrightarrow{L} u_0 \text{ strongly (weakly) in } L^2(\kappa, T; H^1(\Omega_0)),$$

if $\{\|u_n\|_{L^2(\kappa, T; H^1(\Omega_n))}\}$ is bounded and for any $G \Subset \Omega_0$ there is $N(G) > 0$ such that $u_n \in L^2(\kappa, T; H^1(G))$ after $n \geq N$, and $u_n \rightarrow u_0$ strongly in $L^2(\kappa, T; H^1(G))$ ($L^2(\kappa, T; L^2(G))$).

We now combine the weak compactness and lower-semicontinuity with respect to this convergence in the following theorem.

THEOREM 2.6. *Let $\{\Omega_n\}$ ($n=0, 1, \dots$) be a sequence of bounded open sets in R^m with the property G . Let $\Omega_n \xrightarrow{c} \Omega_0$ and the sequence $\{\|u_n\|_{G^\theta(0, T; H^1(\Omega_n))}\}$ be bounded. Then there are $u_0 \in G^\theta(0, T; H^1(\Omega_0))$ and a subsequence of $\{u_n\}$, still denoted as $\{u_n\}$, such that $u_n \xrightarrow{L} u_0$ weakly in $L^2(\theta\delta, T; H^1(\Omega_0))$ for any fixed $\delta > 0$ and*

$$\int_0^T \int_{\Omega_0} h(t, x, u_0, \nabla u_0) dx dt \leq \liminf_{n \rightarrow \infty} \int_0^T \int_{\Omega_n} h(t, x, u_n, \nabla u_n) dx dt,$$

where h is nonnegative and measurable on $R^+ \times R^m \times R \times R^m$. We suppose that for any fixed $G \Subset \Omega_0$ and $t > 0$, the functional $u \rightarrow \int_G h(t, x, u, \nabla u) dx$ is lower-semicontinuous with respect to $L^2(G)$ convergence in $H^1(G)$.

Proof. By Lions' compactness result in Theorem 5.1 of [10] and a similar method used in the proof of Theorem 2.5, we find that $u_0 \in G^\theta(0, T; H^1(\Omega_0))$ and $u_n \xrightarrow{L} u_0$

weakly in $L^2(\theta\delta, T; H^1(\Omega_0))$ for any $\delta > 0$. Let $\{u_{n_k}\}$ be a subsequence of $\{u_n\}$ such that

$$\lim_{k \rightarrow \infty} \int_0^T \int_{\Omega_{n_k}} h(t, x, u_{n_k}, \nabla u_{n_k}) dx dt = \underline{\lim}_{n \rightarrow \infty} \int_0^T \int_{\Omega_n} h(t, x, u_n, \nabla u_n) dx dt.$$

Then for any $G \Subset \Omega_0$ there is a subsequence of $\{u_{n_k}\}$ (depending on G), still denoted as $\{u_{n_k}\}$, such that $u_{n_k}(t) \rightarrow u_0(t)$ strongly in $L^2(G)$ almost everywhere for $t \in (0, T)$. Therefore, for fixed G ,

$$\int_G h(t, x, u_0, \nabla u_0) dx \leq \underline{\lim}_{n \rightarrow \infty} \int_G h(t, x, u_{n_k}, \nabla u_{n_k}) dx \quad \text{a.e. for } t \in (0, T).$$

Thus, as in the proof of Theorem 2.4,

$$\int_G h(t, x, u_0, \nabla u_0) dx \leq \underline{\lim}_{k \rightarrow \infty} \int_{\Omega_{n_k}} h(t, x, u_{n_k}, \nabla u_{n_k}) dx \quad \text{a.e. for } t \in (0, T).$$

Thus,

$$\begin{aligned} \int_0^T \int_G h(t, x, u_0, \nabla u_0) dx dt &\leq \int_0^T \underline{\lim}_{k \rightarrow \infty} \int_{\Omega_{n_k}} h(t, x, u_{n_k}, \nabla u_{n_k}) dx dt \\ &\leq \underline{\lim}_{k \rightarrow \infty} \int_0^T \int_{\Omega_{n_k}} h(t, x, u_{n_k}, \nabla u_{n_k}) dx dt \\ &= \underline{\lim}_{n \rightarrow \infty} \int_0^T \int_{\Omega_n} h(t, x, u_n, \nabla u_n) dx dt, \end{aligned}$$

due to Fatou's lemma. However, G is any open set such that $G \Subset \Omega_0$. Thus we get our conclusion. \square

Remark 2.3. If $h(t, q, s, p) = h_1(t, q, s)$ and h_1 is nonnegative and continuous with respect to s , then h satisfies all the conditions in Theorem 2.6. On the other hand, if $h(t, q, s, p) = h_2(q, s, p)$, the nonnegative function h_2 is continuous on $R^m \times R \times R^m$, and $h_2(q, s, \cdot)$ is strictly convex in R , then from [14] we see that h satisfies all the conditions in Theorem 2.6.

3. Convergences of convex sets. We now study the convergence of closed convex sets in various spaces. Before doing so, we will first examine a further property of our local convergences, which is important to establish existence of optimal shapes.

3.1. Uniformly absolute continuity. We first note a fact about our local convergences. If $u_n \in H^1(\Omega)$ for $n = 0, 1, \dots$ and $u_n \rightarrow u$ in $H^1(\Omega)$ strongly, then it is known that the sequence of integrals $\{\int_{\Omega} [|u_n|^2 + |\nabla u_n|^2] dx\}$ is uniformly absolutely continuous; that is, for any $\varepsilon > 0$ there is $\delta > 0$ such that

$$\int_E (|u_n|^2 + |\nabla u_n|^2) dx \leq \varepsilon, \quad \text{for } n = 1, 2, \dots \text{ provided } \text{mes}(E) \leq \delta.$$

For our local convergences, however, this is not necessarily true.

Example 3.1. Let $\Omega_0 = \Omega_n = (0, 1)$ for $n = 1, 2, \dots$, $u_0 = 0$, and

$$u_n(t) = \begin{cases} -n/t + 1/n, & 0 < t < 1/n^2 \\ 0 & 1/n^2 \leq t < 1 \end{cases}$$

Then $u_n \xrightarrow{L} u_0$ strongly in $H^1(\Omega_0)$. It is true that $\int_0^{1/n^2} (|u_n|^2 + |\nabla u_n|^2) dt \geq 1$, however.

Because of this, $i_n = \int_{\Omega_n} (\nabla v_n, \nabla u_n) dx$ does not necessarily converge to $i_0 = \int_{\Omega_0} (\nabla v_0, \nabla u_0) dx$, even if $u_n \xrightarrow{L} u_0$ and $v_n \xrightarrow{L} v_0$ strongly in $H^1(\Omega_0)$, and this makes it difficult to establish any existence results for optimal shapes. Thus, we now introduce the following result.

DEFINITION 3.1. Let $u_n \in H^1(\Omega_n)$ for $n = 1, 2, \dots$. We say that $\{u_n\}$ is *uniformly absolutely continuous* (U.A.C.) if

$$\int_{E_n} (|u_n|^2 + |\nabla u_n|^2) dx \rightarrow 0,$$

where the measurable sets $E_n \in \Omega_n$ for $n = 1, 2, \dots$, and $\text{mes}(E_n) \rightarrow 0$. We have a similar definition for $u_n \in L^2(\kappa; T; H^1(\Omega_n))$ ($\kappa \geq 0$); $\{u_n\}$ is U.A.C. if $\int_{\kappa}^T \int_{E_n} [|u_n|^2 + |\nabla u_n|^2] dx dt \rightarrow 0$, where E_n satisfies the same conditions. In the same way, we say that $\{u_n\}$ is *weakly uniformly absolutely continuous* if $\int_{E_n} (|u_n|^2 + |\nabla u_n|^2) dx$ (or $\int_{\kappa}^T \int_{E_n} [|u_n|^2 + |\nabla u_n|^2] dx dt$) is replaced by $\int_{E_n} |u_n|^2 dx$ (or $\int_{\kappa}^T \int_{E_n} |u_n|^2 dx dt$) in the above definition.

Example 3.2. If $\Omega_n \subset C$ and there is an extension operator j_n from $H^1(\Omega_n)$ to $H^1(C)$ ($n = 1, 2, \dots$) such that $\sup_n \|j_n\| < \infty$ [5], where C is a bounded open set of R^m with a Lipschitz boundary, then for any $\{u_n\}$ with $u_n \in H^1(\Omega_n)$ for $n = 1, 2, \dots$, which is locally weakly convergent to an element in $H^1(\Omega_0)$, there is $\{u_{n_k}\}$, a subsequence of $\{u_n\}$, such that it is weakly U.A.C. To see this, we note that $j_{n_k} u_{n_k} \rightarrow u_0$ in $L^2(C)$ strongly (see [1]) and $\int_{E_{n_k}} |u_{n_k}|^2 dx \leq \int_{E_{n_k}} |j_{n_k} u_{n_k}|^2 dx \rightarrow 0$, while $\text{mes}(E_{n_k}) \rightarrow 0$. In fact, we can prove that $\{u_n\}$ itself is weakly U.A.C. To do this, we just note that from Sobolev's embedding results [1] (note that C is bounded), there is $\delta > 0$, $\rho > 1$ such that

$$\left(\int_C |j_n u_n|^{2\rho} dx \right)^{1/2\rho} \leq \delta \|j_n u_n\|_{H^1(C)} \leq \delta \sup_n \|j_n\| \|u_n\|_{H^1(\Omega_n)},$$

so that

$$\int_{E_n} |u_n|^2 dx = \int_{E_n} |j_n u_n|^2 dx \leq \left(\int_C |j_n u_n|^{2\rho} dx \right)^{1/\rho} \left(\int_{E_n} dx \right)^{1-1/\rho}.$$

Thus $\{u_n\}$ is weakly U.A.C.

On the other hand, we can prove a similar conclusion for $\{u_n\}$ with $u_n \in G^\theta(0, T; H^1(\Omega_n))$; that is, $\int_0^T \int_{E_n} u_n^2 dx dt \rightarrow 0$ ($E_n \subset \Omega_n$), provided $\{\|u_n\|_{G^\theta(0, T; H^1(\Omega_n))}\}$ is bounded and $\text{mes}(E_n) \rightarrow 0$, because

$$\begin{aligned} \int_0^T \int_{E_n} u_n^2 dx dt &= \int_0^T \int_{E_n} (j_n u_n)^2 dx dt \\ &\leq \int_0^T \int_C \|j_n u_n\|_{L^{2\rho}(C)}^2 dx dt \text{mes}(E_n)^{1-1/\rho} \\ &\leq \delta \sup_n \|j_n\| \text{mes}(E_n)^{1-1/\rho} \int_0^T \int_{\Omega_n} \|u_n\|_{H^1(\Omega_n)}^2 dx dt. \end{aligned}$$

Finally, we note a useful fact: the operator J_n , $(J_n u)(t, x) = (j_n u(t, \circ))(x)$ is also an extension operator from $G^\theta(0, T; H^1(\Omega_n))$ to $G^\theta(0, T; H^1(C))$ for $n = 1, 2, \dots$, and $\{\|J_n\|\}$ is bounded if j_n is also an extension operator from $L^2(\Omega_n)$ to $L^2(C)$ with $\sup_n \|j_n\|_{\mathcal{L}(L^2(\Omega_n), L^2(C))} < \infty$. To see this, we must note that $J_n u$ is strongly measurable to $t > 0$,

$$\int_0^T \|J_n u\|_{H^1(C)}^2 dt \leq \int_0^T \|j_n\| \|u\|_{H^1(\Omega_n)}^2 dt,$$

and

$$\int_0^T \|t^\theta (j_n u_n)'\|_{L^2(C)}^2 dt \leq \int_0^T \|j_n\|_{\mathcal{L}(L^2(\Omega_n), L^2(C))} \|t^\theta u_n'\|_{L^2(\Omega_n)}^2 dt.$$

Remark 3.1. It can be proved by the integral transform formula that if $\{\Omega_n\}$ is bounded in $O^{k,\infty}(k \geq 1)$ and $\{\|u_n\|_{H^1(\Omega_n)}\}$ is bounded, then $\{u_n\}$ is weakly U.A.C.

3.2. Local Mosco convergence. We now give a definition of the convergence of convex sets in some various spaces, similar to the Mosco convergence in [2].

DEFINITION 3.2. Let $\{\Omega_n\}$ ($n = 0, 1, \dots$) be a sequence of bounded open sets in R^m with the property G . Let K_n be a nonempty closed convex set in $H^1(\Omega_n)$ for $n = 0, 1, \dots$. We say that $\{K_n\}$ *locally Mosco converges* to K_0 in $H^1(\Omega_0)$ (denoted as $K_n \xrightarrow{LM} K_0$ in $H^1(\Omega_0)$), if and only if

- (i) $U \in K_0$, provided there is $\{n_k\}$, a subsequence of $\{n\}$, such that $u_{n_k} \in K_{n_k}$ and $u_{n_k} \xrightarrow{L} u$ weakly in $H^1(\Omega_0)$.
- (ii) For any $u \in K_0$, there are $u_n \in K_n$ ($n = 1, 2, \dots$) such that $u_n \xrightarrow{L} u$ strongly in $H^1(\Omega_0)$, and $\{u_n\}$ is U.A.C.

It is clear that even if $\Omega_n = \Omega_0$, $n = 1, 2, \dots$, this concept is not the same as the Mosco convergence of closed convex sets in [2]. To see this, we take $\Omega_n = \Omega_0 = (0, 1)$ and $K_n = K_0 = \{u \in H^1(\Omega_0); u|_{\partial\Omega_0} \geq 0\}$. Then it follows that K_n Mosco converges to K_0 . It is easy to find $u_n \in K_n$ such that $u_n \xrightarrow{L} -1$ weakly in $H^1(\Omega_0)$, which is not in K_0 .

In the following examples, $O_\varepsilon = \{\Omega: \Omega \text{ open}, \Omega \subset C \text{ a bounded open set of } R^m \text{ with a Lipschitz boundary, } \Omega \text{ has the } \varepsilon\text{-cone property for a fixed } \varepsilon > 0\}$, $\Omega_n \in O_\varepsilon$ ($n = 0, 1, \dots$), and $\Omega_n \rightarrow \Omega_0$ in the distance $d_C(\cdot, \cdot)$.

Example 3.3. Let $K_n = H^1(\Omega_n)$ for $n = 0, 1, \dots$. We prove that $K_n \xrightarrow{LM} K_0$ in $H^1(\Omega_0)$. First, if $u_{n_k} \in K_{n_k}$ and $u_{n_k} \xrightarrow{L} u$ weakly in $H^1(\Omega_0)$, then $u \in H^1(\Omega_0)$. Next, for any $u \in H^1(\Omega_0)$, let j_0 be an extension operator from $H^1(\Omega_0)$ to $H^1(C)$ and $\{u_n\} = \{j_0 u|_{\Omega_n}\}$, which is U.A.C. Then, from $G \subset \Omega_n$ ($n \geq N$) for any fixed $G \Subset \Omega_0$ (see § 2), we see that $u_n \xrightarrow{L} u$ strongly in $H^1(\Omega_0)$.

Example 3.4. Let $K_n = H_0^1(\Omega_n)$ for $n = 0, 1, \dots$. If $u_{n_k} \rightarrow u_0$ with $u_{n_k} \in H^1(\Omega_{n_k})$, it follows that $u_0 \in H_0^1(\Omega_0)$ by passing the limits in $(\chi(\Omega_{n_k}) - 1)u_{n_k} = 0$ and by using the fact that $\chi(\Omega_n) \rightarrow \chi(\Omega_0)$ in $L^2(R^m)$ strongly and $\partial\Omega_0$ is Lipschitz, where \underline{u}_n is the zero extension of u_n on R^m . Let $u \in H_0^1(\Omega_0)$. Then there are $v_n \in \mathcal{D}(\Omega_0)$ such that $v_n \rightarrow u$ strongly in $H_0^1(\Omega_0)$. For any v_i , there is an integer $N_i > 0$ such that $v_i \in H_0^1(\Omega_j)$ ($j \geq N_i$) and $N_{i+1} > N_i$ because v_i has a compact support in Ω_0 . We now take $u_j = v_i$ if $N_i \leq j < N_{i+1}$, where $j = N_1, N_j + 1, \dots$. Then we construct $\{u_n\}$ such that $u_n \in H_0^1(\Omega_n)$ and $u_n \xrightarrow{L} u$ strongly in $H_0^1(\Omega_0)$. We see that $K_n \xrightarrow{LM} K_0$ in $H^1(\Omega_0)$ if we can prove that $\{u_n\}$ obtained above is U.A.C. To this end, we must note that $\{u_n\} \subset \mathcal{D}(\Omega_0)$ and $u_n \rightarrow u$ in $H_0^1(\Omega_0)$ implies that $\underline{u}_n \rightarrow \underline{u}$ in $H_0^1(C)$, where \underline{u} is the zero extension of u on C .

Example 3.5. Let $K_n = \{u: u \in H_0^1(\Omega_n), u \geq 0 \text{ almost everywhere in } \Omega_n\}$ for $n = 0, 1, \dots$. We now prove that $K_n \xrightarrow{LM} K_0$ in $H^1(\Omega_0)$. First, if $u_{n_k} \in K_{n_k}$ and $u_{n_k} \xrightarrow{L} u$, then we see that for any $G \Subset \Omega_0$, $u|_G \geq 0$ so that $u \in K_0$. Next, for any $u \in K_0$, from Example 3.4, there are $u_n \in H_0^1(\Omega_n)$ such that $u_n \rightarrow u$ in $H^1(G)$ strongly for any $G \Subset \Omega_0$ and $\{\|u_n\|_{H^1(\Omega_n)}\}$ is bounded. Let $\underline{u}_n = \max(u_n, 0) \in K_n$. Then from [4], we know that $\underline{u}_n \rightarrow u$ strongly in $H^1(G)$ since $u \geq 0$ and that $\|\underline{u}_n\|_{H^1(\Omega_n)} \leq \|u_n\|_{H^1(\Omega_n)}$. On the other hand, from the inequality

$$\int_{E_n} (|\underline{u}_n|^2 + |\nabla \underline{u}_n|^2) dx \leq \int_{E_n} (|u_n|^2 + |\nabla u_n|^2) dx$$

and the fact that $\{u_n\}$ is U.A.C. (see Ex. 3.4), we see that $\{\underline{u}_n\}$ is U.A.C., also.

By the same method we can prove that $K_n \xrightarrow{LM} K_0$, where $K_n = \{u: u \in H^1(\Omega_n), u \geq \varphi \text{ almost everywhere in } \Omega_n, \Phi \in H^1(R^n)\}$ or $K_n = \{u: u \in H^1(\Omega_n), |u| \leq 1 \text{ almost everywhere in } \Omega_n\}$ for $n = 0, 1, \dots$.

Example 3.6. Let $K_n = \{u: u \in H^1(\Omega_n), \int_{\Omega_n} |\nabla u| dx \leq 1\}$ for $n = 0, 1, 2, \dots$. We prove here that $K_n \xrightarrow{LM} K_0$. First, if $u_{n_k} \xrightarrow{L} u$ weakly in $H^1(\Omega_0)$, then for any $G \Subset \Omega_0$, $u_{n_k} \rightarrow u$ weakly in $H^1(G)$. Thus, $\int_G |\nabla u| dx \leq 1$; that is, $u \in K_0$. Next, for any $u \in K_0$, let $\underline{u} = j_0 u$, where j_0 is an extension operator from $H^1(\Omega_0)$ to $H^1(C)$. We then have

$$\int_{\Omega_n} |\nabla \underline{u}| dx \leq \int_{\Omega_0} |\nabla u| dx + \int_{\Omega_n \setminus (\Omega_n \cap \Omega_0)} |\nabla \underline{u}| dx \leq 1 + \varepsilon_n,$$

where $\varepsilon_n \rightarrow 0$ because $\chi_{\Omega_n} \rightarrow \chi_{\Omega_0}$ strongly in $L^2(R^m)$, as seen before. Let $u_n = \underline{u}_n / 1 + \varepsilon_n \in K_n$, where $\underline{u}_n = \underline{u}|_{\Omega_n}$. Then, we have that $u_n \rightarrow u$ strongly in $H^1(G)$ for any $G \Subset \Omega_0$, and $\{u_n\}$ is U.A.C.

In a similar method, we show that $K_n \xrightarrow{LM} K_0$ in $H^1(\Omega_0)$, where $K_n = \{u: u \in H_0^1(\Omega_n), |\nabla u| \leq 1 \text{ almost everywhere in } \Omega_n\}$ for $n = 0, 1, \dots$.

Remark 3.2. It can easily be proved that the conclusions for Examples 3.3–3.6 still hold if we replace the condition $\{\Omega_n\} \subset O_\varepsilon$ and $\Omega_n \rightarrow \Omega_0$ in the distance d_C , by $\{\Omega_n\} \subset O^{k,\infty}$ and $\Omega_n \rightarrow \Omega_0$ in $O^{k,\infty}$ ($k \geq 1$). We can also prove the conclusions for some other classes of open sets by the fact that $\Omega_n \xrightarrow{L} \Omega_0$ and $\chi_{\Omega_n} \rightarrow \chi_{\Omega_0}$ as $\Omega_n \rightarrow \Omega_0$ (see Remark 2.2).

We now study the evolution case. First, we give the following definition.

DEFINITION 3.3. Let $\{\Omega_n\}$ ($n = 0, 1, \dots$) be a sequence of bounded open sets in R^m with the property G . Let K_n be a nonempty close convex set in the space $L^2(0, T; H^1(\Omega_n))$ for $n = 0, 1, \dots$. We say that $\{K_n\}$ *locally Mosco converges* to K_0 in $L^2(0, T; H^1(\Omega_0))$ (to be denoted as $K_n \xrightarrow{LM} K_0$ in $L^2(0, T; H^1(\Omega_0))$) if

(i) $u \in K_0$ provided there is $\{u_{n_k}\}$, where $\{n_k\}$ is a subsequence of $\{n\}$, such that $u_{n_k} \in K_{n_k}$ and $u_{n_k} \xrightarrow{L} u$ weakly in $L^2(0, T; H^1(\Omega_0))$.

(ii) For any $u \in K_0$, there are $u_n \in K_n$ ($n = 1, 2, \dots$) such that $u_n \xrightarrow{L} u$ strongly in $L^2(0, T; H^1(\Omega_0))$ and $\{u_n\}$ is U.A.C.

Example 3.7. Let $K_n = L^2(0, T; H^1(\Omega_n))$ for $n = 0, 1, \dots$. As in Example 3.3, we can prove that $K_n \xrightarrow{LM} K_0$ in $L^2(0, T; H^1(\Omega_0))$ if we note that there is an extension operator $J: L^2(0, T; H^1(\Omega_0))$ to $L^2(0, T; H^1(R^n))$. To see this, let j_0 be an extension operator from $H^1(\Omega_0)$ to $H^1(C)$ and $(J_0 u)(t, x) = [j_0 u(t, x)](x)$. Then, J_0 is one from $L^2(0, T; H^1(\Omega_0))$ to $L^2(0, T; H^1(R^m))$, as in Example 3.2. We can also prove a similar conclusion for $K_n = L^2(0, T; H_0^1(\Omega_n))$ for $n = 0, 1, 2, \dots$ if we note that $\mathcal{D}((0, T) \times \Omega_0)$ is dense in $L^2(0, T; H_0^1(\Omega_0))$ and $(0, T) \times \Omega_n \xrightarrow{(0, T) \times C} (0, T) \times \Omega_0$ as $\Omega_n \xrightarrow{C} \Omega_0$.

Example 3.8. Let $K_n = \{u: u \in L^2(0, T; H^1(\Omega_n)), u \geq \psi \text{ almost everywhere, where } \psi \text{ is in } L^2(0, T; H^1(R^n))\}$. We now prove that $K_n \xrightarrow{LM} K_0$ in $L^2(0, T; H^1(\Omega_0))$. If $u_n \in K_n$ and $u_n \xrightarrow{L} u$ weakly in $L^2(0, T; H^1(\Omega_0))$, then for any $G \Subset \Omega_0$, $u_n \rightarrow u$ weakly in $L^2(0, T; L^2(G))$. Thus $u \geq \psi$ almost everywhere in $(0, T) \times \Omega_0$; that is, $u \in K_0$. Next, for any $u \in K_0$, let $u_n = \max(J_0 u, \psi)|_{\Omega_n} \in K_n$ for $n = 1, 2, \dots$. Then, for any $G \Subset \Omega_0$, there is $N > 0$ such that $u_n \in L^2(0, T; H^1(G))$ ($n \geq N(G)$) and $u_n \rightarrow u$ strongly in $L^2(0, T; H^1(G))$ because there is $N(G) > 0$ such that $G \subset \Omega_n$ ($n \geq N(G)$). It is clear that $\{u_n\}$ is U.A.C.

In the same way, we can also prove that $K_n \xrightarrow{LM} K_0$, where $K_n = \{u: u \in L^2(0, T; H_0^1(\Omega_n)), u \geq \psi, \psi \leq 0, \psi \in L^2(0, T; H^1(R^m))\}$.

Remark 3.3. For Examples 3.7 and 3.8, we have the same remark as Remark 3.2.

4. Existence of optimal shape design for variational inequalities. In this section, \mathcal{B} will be a class of bounded open sets in R^m , which are contained in a fixed bounded open set. We further suppose that there is an open set C of R^m and a topology τ on

\mathcal{B} , which is finer than the topology given by the distance d_C , such that (\mathcal{B}, τ) is compact in the sense that every sequence in \mathcal{B} has a convergent subsequence. Let K_Ω be a nonempty closed convex set in $H^1(\Omega)$ for any $\Omega \in \mathcal{B}$. For any $u, v \in H^1(\Omega)$, let

$$a(u, v) = \int_{\Omega} [(A(x)\nabla u(x), \nabla v(x))_{R^m} + a(x)u(x)v(x)] dx,$$

$$(f, u) = \int_{\Omega} f(x)u(x) dx,$$

where $A \in [L^\infty(R^m)]^{m \times m}$ is a positive matrix, $a \in L^\infty(R^m)$, $a(x) \geq 0$, and $f \in L^2(R^m)$. We suppose that $a(u, u)$ is uniformly coercive; that is, $a(u - v, u - v) \geq c\|u - v\|_{H^1(\Omega)}^2$ for any $u, v \in K_\Omega$, where the constant c is independent of Ω .

4.1. Existence of optimal shape design for elliptic variational inequalities. We now consider the following optimal shape design problem (OSD):

$$\min_{\Omega \in \mathcal{B}} \int_{\Omega} g(x, u, \nabla u) dx,$$

subject to

$$(4.1) \quad a(u, v - u) \geq (f, v - u), \quad u \in K_\Omega, \text{ for every } v \in K_\Omega,$$

where g is a nonnegative measurable function on $R^m \times R \times R^m$, $g(x, t, \cdot)$ is convex on R^m , and $g(x, \cdot, \cdot)$ is continuous on $R \times R^m$.

We now give a general existence result for the optimal shapes.

THEOREM 4.1. *If $K_{\Omega_n} \xrightarrow{LM} K_{\Omega_0}$ when $\Omega_n \rightarrow \Omega_0$ in the topology τ , then (OSD) has at least one solution.*

Proof. Let $\{\Omega_n\}$ ($n = 1, 2, \dots$) be a minimizing sequence of (OSD) and u_n be the solution of (4.1) on Ω_n . From the condition, we know that there is a subsequence of $\{\Omega_n\}$, still denoted as $\{\Omega_n\}$, and $\Omega_0 \in \mathcal{B}$ such that $\Omega_n \rightarrow \Omega_0$ in the topology τ . Thus $\Omega_n \xrightarrow{C} \Omega_0$ and $\chi_{\Omega_n} \rightarrow \chi_{\Omega_0}$ strongly in $L^1(R^m)$ (so in $L^2(R^m)$). From

$$(4.2) \quad a(u_n, v - u_n) \geq (f, v - u_n), \quad u_n \in K_{\Omega_n}, \text{ for all } v \in K_{\Omega_n},$$

Schwartz's inequality, and the uniform coercivity of $a(\cdot, \cdot)$, we see that there is $M > 0$ such that

$$c\|u_n - v\|_{H^1(\Omega)}^2 \leq a(u_n - v, u_n - v) \leq (M\|v\|_{H^1(\Omega)} + \|f\|_{L^2(R^m)})\|u_n - v\|_{H^1(\Omega)},$$

for any $v \in K_{\Omega_n}$. Let v_0 be a fixed element in K_{Ω_0} and $v_n \xrightarrow{L} v_0$ strongly in $H^1(\Omega_0)$ with $v_n \in K_{\Omega_n}$. We see that $\{\|v_n\|_{H^1(\Omega_n)}\}$ is bounded and thus, by letting $v = v_n$ for $n = 1, 2, \dots$ in the inequalities above, it follows that $\{\|u_n\|_{H^1(\Omega_n)}\}$ is also bounded. By Theorem 2.5, we know that there are $u_0 \in H^1(\Omega_0)$ and a subsequence of $\{u_n\}$, still denoted as $\{u_n\}$, such that $u_n \xrightarrow{L} u_0$ weakly in $H^1(\Omega_0)$. We now show that u_0 is the solution of (4.1) on Ω_0 . To this end, we first note that $u_0 \in K_{\Omega_0}$ because $K_{\Omega_n} \xrightarrow{LM} K_{\Omega_0}$. For the same reason, for any $v \in K_{\Omega_0}$, there are $v_n \in K_{\Omega_n}$ such that $v_n \xrightarrow{L} v$ strongly in $H^1(\Omega_0)$. Next, for any $\varepsilon > 0$ there are $G \Subset \Omega_0$ and $N > 0$ such that $G \Subset \Omega_n$ and $\text{mes}(\Omega_0 \setminus G) + \text{mes}(\Omega_n \setminus G) \leq \varepsilon$ for $n \geq N$ because $\chi_{\Omega_n} \rightarrow \chi_{\Omega_0}$ strongly in $L^2(R^m)$. For such fixed $G \Subset \Omega_0$,

$$(4.3) \quad \lim_{n \rightarrow \infty} \int_G f(u_n - v_n) dx = \int_G (f, u_0 - v) dx,$$

$$(4.4) \quad \lim_{n \rightarrow \infty} \int_G [(A\nabla u_n, \nabla v_n)_{R^m} + au_nv_n] dx = \int_G [(A\nabla u_0, \nabla v)_{R^m} + au_0v] dx,$$

for any $v \in K_{\Omega_0}$, where $u_n \xrightarrow{L} u_0$ weakly, $v_n \xrightarrow{L} v$ strongly in $H^1(\Omega_0)$, and $\{v_n\}$ is U.A.C. We now estimate $j_1 = |\int_{\Omega_n \setminus G} [(A \nabla u_n, \nabla v_n) a u_n v_n] dx|$ and $j_2 = |\int_{\Omega_n \setminus G} f(u_n - v_n) dx|$. We see that j_2 can be arbitrary small since

$$j_2 \leq \|f\|_{L^2(\Omega_n \setminus G)} (\|u_n\|_{L^2(\Omega_n \setminus G)} + \|v_n\|_{L^2(\Omega_n \setminus G)}).$$

From the fact that $\{v_n\}$ is U.A.C. and from the same reasons as above, we also know that j_1 can be arbitrarily small from Schwartz's inequality. Combining (4.3) and (4.4), we get

$$(4.5) \quad \int_{\Omega_0} f(v - u_0) dx = \lim_{n \rightarrow \infty} \int_{\Omega_n} f(v_n - u_n) dx,$$

$$(4.6) \quad \int_{\Omega_0} [(A \nabla u_0, \nabla v) + a u_0 v] dx = \lim_{n \rightarrow \infty} \int_{\Omega_n} [(A \nabla u_n, \nabla v_n) + a u_n v_n] dx.$$

On the other hand, from Corollary 2.1,

$$\int_{\Omega_0} [(A \nabla u_0, \nabla u_0) + a u_0 u_0] dx \leq \lim_{n \rightarrow \infty} \int_{\Omega_n} [(A \nabla u_n, \nabla u_n) + a u_n u_n] dx.$$

Together with (4.5) and (4.6), we infer that, for any $v \in K_{\Omega_0}$,

$$\begin{aligned} (f, u_0 - v) &= \int_{\Omega_0} f(u_0 - v) dx = \lim_{n \rightarrow \infty} \int_{\Omega_n} f(u_n - v_n) dx \\ &\geq \lim_{n \rightarrow \infty} a(u_n, u_n - v_n) \geq a(u_0, u_0 - v). \end{aligned}$$

That is, u_0 is the solution of (4.1) on Ω_0 . To see that Ω_0 is an optimal shape, we only need to use Corollary 2.2, that $u_n \xrightarrow{L} u_0$ weakly in $H^1(\Omega_0)$, and that $\{\Omega_n\}$ is a minimizing sequence of (OSD). \square

COROLLARY 4.1. *Let $\mathcal{B} = O_\varepsilon$ and $\{K_\Omega\}$ in (OSD) be the same as one in Examples 3.3–3.6. Then (OSD) has at least one solution.*

Remark 4.1. If \mathcal{B} is a closed bounded set in $O^{k,\infty}$ ($k \geq 2$) (see § 2) in (OSD), Theorem 4.1 and Corollary 4.1 are still true, from Remarks 2.2. and 3.2.

Remark 4.2. Theorem 4.1 is a very general existence result, and it can be used in some other classes of open sets. Furthermore, we see from the theorem that in many cases (see, e.g., Example 3.3) we still have existence results without assuming the condition that there are uniform extension operators (Theorem 2.3), which was assumed in other works [5]–[7], [13].

4.2. Existence of optimal shape design for evolution variational inequalities. Let \mathcal{B} have the same meaning as above and K_Ω be a nonempty closed convex set in $L^2(0, T; H^1(\Omega))$ for every $\Omega \in \mathcal{B}$. We consider the following evolution optimal shape design problem:

$$(OSE) \quad \min_{\Omega \in \mathcal{B}} \int_0^T \int_{\Omega} h(t, x, u, \nabla u) dx dt,$$

subject to

$$(4.7) \quad \int_0^T (u', v - u) dt + \int_0^T a(u, v - u) dt \geq \int_0^T (f, v - u) dt,$$

$$u(0, x) = \psi(x), \quad u \in K_\Omega \cap W(0, T; H^1(\Omega)) \text{ for any } v \in K_\Omega,$$

where $\psi \in L^2(R^m)$, $f \in L^2(R^{m+1})$, $(u', v) = \langle u', v \rangle_{(H^1(\Omega))^* \times H^1(\Omega)}$, and h satisfies conditions in Theorem 2.6. Furthermore, we suppose that $A(\cdot)$ and $a(\cdot)$ in $a(u, v)$ are continuous on R^m . Thus, we give the following result.

THEOREM 4.2. *If (4.7) has a unique solution for any $\Omega \in \mathcal{B}$ and the solution u_Ω is in $G^0(0, T; H^1(\Omega))$ such that the following uniform boundedness condition holds:*

$$\|u_\Omega\|_{G^0(0, T; H^1(\Omega))} \leq d, \quad \text{where } d \text{ is independent of } \Omega \in \mathcal{B},$$

then (OSE) has at least one solution if $K_{\Omega_n} \xrightarrow{LM} K_{\Omega_0}$ in $L^2(0, T; H^1(\Omega_0))$ as $\Omega_n \rightarrow \Omega_0$ in (\mathcal{B}, τ) .

Proof. We use the same method as in the proof of Theorem 4.1. Let $\{\Omega_n\}$ be a minimizing sequence of (OSE). From Theorem 2.6 and the uniform boundedness condition, we know that there is a subsequence of $\{u_n\}$, still denoted as $\{u_n\}$, and $u_0 \in G^0(0, T; H^1(\Omega_0))$ such that $u_n \xrightarrow{L} u_0$ weakly in $L^2(0, T; H^1(\Omega_0))$, where u_n is the solution of (4.7) on Ω_n . We only need to prove that u_0 is the solution of (4.7) on Ω_0 due to Theorem 2.6. First, by the same method in the proof of Theorem 4.1, we have, for any $v \in K_{\Omega_0}$,

$$(4.8) \quad \int_0^T \int_{\Omega_0} f(u_0 - v) \, dx \, dt = \lim_{n \rightarrow \infty} \int_0^T \int_{\Omega_n} f(u_n - v_n) \, dx \, dt,$$

where $\{v_n\}$ is U.A.C., $v_n \xrightarrow{L} v$ strongly in $L^2(0, T; H^1(\Omega_0))$. Similarly, from Theorem 2.6, Remark 2.1, and Example 3.2,

$$(4.9) \quad \begin{aligned} & \int_0^T \int_{\Omega_0} [(A \nabla u_0, \nabla(u_0 - v)) + a u_0(u_0 - v)] \, dx \, dt \\ & \leq \lim_{n \rightarrow \infty} \int_0^T \int_{\Omega_n} [(A \nabla u_n, \nabla(u_n - v_n)) + a u_n(u_n - v_n)] \, dx \, dt. \end{aligned}$$

We now prove that

$$(4.10) \quad \int_0^T (u'_0, v) \, dt = \lim_{n \rightarrow \infty} \int_0^T (u'_n, v_n) \, dt.$$

To see this, we first note that for any $G \Subset \Omega_0$,

$$\int_0^T \int_G u'_0 v \, dx \, dt = \lim_{n \rightarrow \infty} \int_0^T \int_G u'_n v_n \, dx \, dt,$$

since $u'_n \rightarrow u'_0$ weakly in $L^2(0, T; L^2(G))$ from the fact that $u_n \rightarrow u_0$ in $L^2(0, T; L^2(G))$ strongly and that $\{\|u'_n\|_{L^2(0, T; L^2(G))}\}$ is bounded. Next, we note that for any $\varepsilon > 0$ there are $G \Subset \Omega_0$ and $N > 0$ such that $G \subset \Omega_n$ ($n \geq N$) and $\text{mes}(\Omega_n \setminus G) + \text{mes}(\Omega_0 \setminus G) \leq \varepsilon$ such that

$$\begin{aligned} & \left| \int_0^T \int_{\Omega_n \setminus G} u'_n v_n \, dx \, dt \right| + \left| \int_0^T \int_{\Omega_0 \setminus G} u'_0 v \, dx \, dt \right| \\ & \leq \|u'_n\|_{L^2(0, T; L^2(\Omega_n))} \|v_n\|_{L^2(0, T; H^1(\Omega_n \setminus G))} + \|u'_0\|_{L^2(0, T; L^2(\Omega_0))} \|v\|_{L^2(0, T; H^1(\Omega_0 \setminus G))}, \end{aligned}$$

which can be arbitrarily small if $\text{mes}(\Omega_0 \setminus G)$ is small enough, because $\{v_n\}$ is U.A.C. and $\{\|u_n\|_{G^0(0, T; H^1(\Omega))}\}$ is bounded. So we get (4.10) from

$$\int_0^T (u'_n, v_n) \, dt = \int_0^T \int_{\Omega_n} u'_n v_n \, dx \, dt.$$

Finally, we prove that

$$(4.11) \quad \int_0^T (u'_0, u_0) \, dt \leq \lim_{n \rightarrow \infty} \int_0^T (u'_n, u_n) \, dt.$$

To do this, we note that since $u_n \rightarrow u_0$ strongly in $L^2(0, T; L^2(G))$ for any $G \Subset \Omega_0$ and $\{\|u_n\|_{G^0(0, T; H^1(\Omega_n))}\}$ is bounded, then it follows from Lions' compactness embedding result in [10] that $u_n \rightarrow u_0$ weakly in $W(0, T)$ for any $G \Subset \Omega_0$. On the other hand, $W(0, T)$ is embedded in $C(0, T; L^2(G))$ [9], so $u_n(T, x) \rightharpoonup u_0(T, x)$ weakly in $L^2(\Omega_0)$. Thus (see the proof of Theorem 2.4),

$$\begin{aligned} 2 \int_0^T (u'_0, u_0) dt &= \int_{\Omega_0} u_0^2(T, x) dx - \int_{\Omega_0} \psi^2(x) dx \\ &\leq \liminf_{n \rightarrow \infty} \int_{\Omega_n} u_n^2(T, x) dx - \lim_{n \rightarrow \infty} \int_{\Omega_n} \psi^2(x) dx \\ &= \liminf_{n \rightarrow \infty} \int_{\Omega_n} [u_n^2(T, x) - \psi^2(x)] dx = 2 \liminf_{n \rightarrow \infty} \int_0^T (u'_n, u_n) dt. \end{aligned}$$

Combining (4.8)–(4.11), we infer that u_0 is the solution of (4.7) on Ω_0 . Thus Ω_0 is an optimal shape from Theorem 2.6. \square

COROLLARY 4.2. *Let $\psi = 0$, $f \in C(0, T; H^{-1}(R^n)) \cap L^2(R^{m+1})$ and $f(0) = 0$ in (OSE). Suppose for every fixed $\Omega \in \mathcal{B}$, that there is a constant $\rho > 0$, independent of Ω , such that, for any $v \in K_\Omega$ and $s \geq 0$,*

$$G(s)v + G^*(s)v - G^*(s)G(s)v + (\rho - 1)v \in \rho K_\Omega,$$

where $G(s)$ is the semigroup introduced in [10, p. 294]. Then (OSE) has at least one solution, provided $K_{\Omega_n} \xrightarrow{LM} K_{\Omega_0}$ as $\Omega_n \xrightarrow{\tau} \Omega_0$.

Proof. We only need to note that the uniform boundedness condition holds from the proof of Theorem 9 in [10, p. 294]. \square

Remark 4.3. We have the same remark for Theorem 4.2 and Corollary 4.2 as Remark 4.1.

Another important case where the uniform boundedness condition holds can be found in [3, p. 124]. However, this condition is somewhat restrictive. In some important cases, we can replace it by the common condition that $\{\|u_\Omega\|_{G^0(0, T; H^1(\Omega))}\}$ is bounded.

Before giving further results in this direction, we first give a definition.

DEFINITION 4.1. Let K_Ω be a closed convex set in $H^1(\Omega)$ for any $\Omega \in \mathcal{B}$. We say that $\{K_\Omega\}(\Omega \in \mathcal{B})$ is L^2 -locally closed if for any $\{\Omega_n\} \subset \mathcal{B}$ ($n = 0, 1, \dots$) such that $\Omega_n \xrightarrow{\subseteq} \Omega_0$, we have $u \in K_{\Omega_0}$, provided there are $u_{n_k} \in K_{\Omega_{n_k}}$ satisfying that for any $G \Subset \Omega_0$, there is $N(G) > 0$ such that $u_{n_k} \in L^2(G)$ ($k \geq N$) and $u_{n_k} \rightarrow u$ strongly in $L^2(G)$ (in Definition 3.2(i), we have an additional requirement that $\{\|u_{n_k}\|_{H^1(\Omega_{n_k})}\}$ be bounded).

We note that all $\{K_\Omega\}$ in Examples 3.3–3.6 are L^2 -locally closed.

In the following, $\mathcal{B} = \mathcal{O}_\varepsilon$ and the condition that $u \in K_\Omega \cap W(0, T)$ in (OSE) will be replaced by $u \in K_\Omega \cap G^{1/2}(0, T; H^1(\Omega))$, and we further require that $a(u, v) = a(v, u)$ for any $u, v \in H^1(\Omega)$ to ensure that we can use the existence theorem of solutions for evolution variational inequalities in [3].

THEOREM 4.3. *In (OSE), let $\mathcal{B} = \mathcal{O}_\varepsilon$ and $K_\Omega = \{v \in L^2(0, T; H^1(\Omega)); v(t) \in \underline{K}_\Omega \text{ almost everywhere for } t \in (0, T)\}$, where \underline{K}_Ω is a closed convex set in $H^1(\Omega)$ and $\psi|_\Omega \in \underline{K}_\Omega$ for every $\Omega \in \mathcal{O}_\varepsilon$. Then (OSE) has at least one solution, provided $\{\underline{K}_\Omega\}$ is L^2 -locally closed and $\underline{K}_{\Omega_n} \xrightarrow{LM} \underline{K}_{\Omega_0}$ in $H^1(\Omega_0)$ as $\Omega_n \rightarrow \Omega_0$ in \mathcal{O}_ε .*

Proof. First we note that (OSE) now is the same as

$$(4.12) \quad \min_{\Omega \in \mathcal{O}_\varepsilon} \int_0^T \int_\Omega h(t, x, u, \nabla u) dx dt, \\ \int_0^T [(u'(t), v - u(t)) + a(u(t), v - u(t))] \rho(t) dt \geq \int_0^T (f, v - u(t)) \rho(t) dt,$$

where $u(t) \in \underline{K}_\Omega$ almost everywhere, $u \in G^{1/2}(0, T; H^1(\Omega))$, for any $v \in \underline{K}_\Omega$ and $\rho \in \mathcal{D}(0, T)$, $\rho \geq 0$. From [3] we see that for any $\Omega \in O_\varepsilon$, (4.12) has a unique solution in $G^{1/2}(0, T; H^1(\Omega))$ and that $\{\|u_\Omega\|_{G^{1/2}(0, T; H^1(\Omega))}\}$ is bounded. Suppose $\{\Omega_n\}$ is a minimizing sequence of (OSE) and $\Omega_n \rightarrow \Omega_0 \in O_\varepsilon$. Then we find a subsequence of $\{u_n\}$, still denoted as $\{u_n\}$, and a function u_0 in $G^{1/2}(0, T; H^1(\Omega_0))$ such that $u_n \xrightarrow{L} u_0$ weakly in $L^2(\delta, T; H^1(\Omega_0))$ for any $\delta > 0$, where u_n is the solution of (4.12) on Ω_n for $n = 1, 2, \dots$. Thus for any $G \Subset \Omega_0$, there is a subsequence of $\{u_n\}$, still denoted as $\{u_n\}$, such that $u_n(t) \rightarrow u_0(t)$ strongly in $L^2(G)$ almost everywhere for $t \in (0, T)$. Therefore, as in the proof of Theorem 2.5 in § 2, there is $\{u_{n_k}(t)\}$ ($k = 1, 2, \dots, n$), a subsequence of $\{u_n(t)\}$, such that $u_{n_k}(t) \in \underline{K}_{\Omega_k}$ almost everywhere for $t \in (0, T)$, and for any fixed $G \Subset \Omega_0$, $u_{n_k}(t) \rightarrow u_0(t)$ strongly in $L^2(G)$ almost everywhere for $t \in (0, T)$. Thus $u_0(t) \in \underline{K}_{\Omega_0}$ almost everywhere for $t \in (0, T)$ because of the L^2 -closeness assumption of $\{\underline{K}_\Omega\}$; that is, $u_0 \in K_{\Omega_0}$. We prove that u_0 is the solution of (4.12) on Ω_0 . First, for any $v \in \underline{K}_{\Omega_0}$, there is $v_n \in \underline{K}_{\Omega_n}$ for $n = 1, 2, \dots$ such that $v_n \xrightarrow{L} v$ strongly in $H^1(\Omega_0)$ (thus, $v_n \xrightarrow{L} v$ strongly in $L^2(0, T; H^1(\Omega_0))$) and that $\{v_n\}$ is U.A.C. Next, from the fact that $\{\|u_n\|_{G^{1/2}(0, T; H^1(\Omega_n))}\}$ is bounded, we infer that for any $\delta > 0$, $\{\|u_n\|_{G^0(\delta, T; H^1(\Omega_n))}\}$ is bounded. So, as in the proof Theorem 4.2, we have, from the fact that $\rho \geq 0$ and that ρ has a compact support in $(0, T)$,

$$(4.13) \quad \int_0^T (f, v - u_0) \rho \, dt = \lim_{n \rightarrow \infty} \int_0^T (f, v_n - u_n) \rho \, dt,$$

$$(4.14) \quad \int_0^T a(u_0, u_0 - v) \rho \, dt \leq \varliminf_{n \rightarrow \infty} \int_0^T a(u_n, u_n - v_n) \rho \, dt,$$

for any fixed $v \in \underline{K}_{\Omega_0}$. Next, note $\rho \in \mathcal{D}(0, T)$. Thus we have for any fixed $G \Subset \Omega_0$

$$\begin{aligned} 2 \int_0^T (u'_0, u_0)_{L^2(G)} \rho \, dt &= - \int_0^T \rho' \|u_0\|_{L^2(G)}^2 \, dt \\ &= - \lim_{n \rightarrow \infty} \int_0^T \rho' \|u_n\|_{L^2(G)}^2 \, dt \\ &= \lim_{n \rightarrow \infty} 2 \int_0^T (u'_n, u_n)_{L^2(G)} \rho \, dt, \end{aligned}$$

because $u_n \xrightarrow{L} u_0$ weakly in $L^2(\delta, T; H^1(\Omega_0))$ for any $\delta > 0$. On the other hand, we have uniformly bounded extension operators j_n from $H^1(\Omega_n)$ to $H^1(C)$ (see § 2). From Example 3.2, we know, as $\text{mes}(\Omega_0 \setminus G) \rightarrow 0$,

$$\left| \int_\delta^T \int_{\Omega_0 \setminus G} \rho u'_0 u_0 \, dx \, dt \right| \quad \text{and} \quad \left| \int_\delta^T \int_{\Omega_n \setminus G} \rho u'_n u_n \, dx \, dt \right|$$

can be arbitrarily small for fixed ρ because

$$2 \left| \int_\delta^T \int_{\Omega_n \setminus G} \rho u'_n u_n \, dx \, dt \right| = \left| \int_\delta^T \int_{\Omega_n \setminus G} \rho' u_n^2 \, dx \, dt \right|$$

for a $\delta > 0$. Thus, we infer that

$$(4.15) \quad \int_0^T (u'_0, u_0) \rho \, dt \leq \varliminf_{n \rightarrow \infty} \int_0^T (u'_n, u_n) \rho \, dt.$$

It only remains to prove that

$$(4.16) \quad \int_0^T (u'_0, v) \rho \, dt = \lim_{n \rightarrow \infty} \int_0^T (u'_n, v_n) \rho \, dt.$$

To show this, we note that for any $\delta > 0$, $\{\int_{\delta}^T \|u'_n\|_{L^2(\Omega_n)}^2 dt\}$ is bounded. Thus, $\int_{\delta}^T (u'_0, v) \rho dt = \lim_{n \rightarrow \infty} \int_{\delta}^T (u'_n, v_n) \rho dt$, as before. On the other hand, ρ has a compact support in $(0, T)$, so we get (4.16). We see now that $u_0 \in G^{1/2}(0, T; H^1(\Omega_0))$, and u_0 satisfies (4.12). Thus we know that Ω_0 is an optimal shape from Theorem 2.6. \square

Combining Theorem 4.3 and Examples 3.7 and 3.8, we can easily give many examples about the existence of optimal shapes in the evolution case.

5. Discussion. It seems that we cannot deal with such cases as $K_{\Omega} = \{v \in H^1(\Omega), v|_{\partial\Omega} \geq 0\}$. However, we must realize that the local convergences given in this paper are only some special cases of a general local convergence theory. In further papers we will develop a general local convergence theory, which seems very useful in the studies of domain optimization. We will see that by using some suitable local convergences, we can treat much wider ranges of domain optimization problems.

Acknowledgments. The authors thank the referees for their valuable comments and suggestions.

REFERENCES

- [1] A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
- [2] H. ATTOUCH, *Variational Convergence for Functions and Operators*, Pitman, London, 1984.
- [3] V. BARBU, *Optimal Control of Variational Inequalities*, Res. Notes Math., 100, Pitman, London, 1984.
- [4] C. BAIocchi AND A. CAPELO, *Variational and Quasi-Variational Inequalities*, John Wiley, New York, 1984, p. 99.
- [5] D. CHENAIS, *On the existence of a solution in a domain identification problem*, J. Math. Anal. Appl., 52 (1975), pp. 189–219.
- [6] N. FUJII, *Lower semicontinuity in domain optimization problems*, JOTA, 58 (1988), pp. 407–421.
- [7] J. HASLINGER AND P. NEITTAANMAKI, *On the design of the optimal covering of an obstacle*, Lecture Notes Control Inform. Sci., 100, J. P. Zolesio, ed., Springer-Verlag, Berlin, New York, 1987, pp. 153–191.
- [8] ———, *On the existence of optimal shapes in contact problems*, Comput. Mech., 1 (1986), pp. 293–299.
- [9] J. L. LIONS, *Non-homogeneous Boundary Value Problems and Applications (I)*, Springer-Verlag, New York, 1972.
- [10] ———, *Quelques Méthodes de Résolution des Problèmes aux Limites Nonlinéaires*, Dunod and Gauthier-Villars, Paris, 1969.
- [11] F. MURAT AND J. SIMON, *Optimal control with respect to domains*, Ph.D. thesis, University of Paris VI, 1985. (In French.)
- [12] ———, *Studies on optimal shape design problems*, Lecture Notes Comput. Sci., 41, J. Cea, ed., Springer-Verlag, Berlin, 1976.
- [13] J. PIRONNEAU, *Optimal Shape Design for Elliptic Systems*, Springer-Verlag, New York, 1984.
- [14] J. SERRIN, *On the definition and the properties of certain variational integrals*, Trans. Amer. Math. Soc., 101 (1961), pp. 139–161.

STABLE SOLUTIONS OF REAL ALGEBRAIC MATRIX RICCATI EQUATIONS*

ANDRÉ C. M. RAN† AND LEIBA RODMAN‡

Abstract. Various stability properties of real symmetric solutions of algebraic matrix Riccati equations with real coefficients are studied. The stability is understood in the sense of robustness, i.e., a solution is stable if, roughly speaking, every nearby equation has a nearby solution.

Key words. Riccati equations, stability, Lipschitz-stability

AMS(MOS) subject classifications. 15A24, 93C35

1. Introduction. We consider the algebraic matrix Riccati equations

$$(1.1) \quad XDC + XA + A^T X - C = 0,$$

where A , D , and C are given $n \times n$ real matrices with $C = C^T$, $D = D^T$, and $D \geq 0$ (we use the notation $X \geq Y$ for real symmetric matrices X and Y to indicate that the difference $X - Y$ is positive semidefinite; the superscript “ T ” stands for the transposed matrix). Only real symmetric solutions X of (1.1) will be considered in this paper. Such solutions, and especially the maximal solution, play a crucial role in the classical quadratic control problems in continuous time (see, e.g., [KS], [Br], [K], or practically any book on linear control systems). In recent years, many new important applications of (1.1) have appeared (we mention only one of them here: H^∞ control; see, e.g., [GGLD], [ZK], and [BC]), and numerical algorithms for finding real symmetric solutions of the algebraic matrix Riccati equation are being developed (see, e.g., [BM], [By], and [L]). Also, real symmetric solutions that are not maximal appear in the optimal feedback control for certain linear quadratic problems (see [T] and [ST]).

In this paper we characterize real symmetric solutions X_0 of (1.1) with various stability properties. We say that a solution X_0 is stable if, roughly speaking, any equation with slightly perturbed coefficients A' , D' , and C' (the perturbed coefficients D' and C' must satisfy the same symmetry conditions as (1.1), i.e., $C' = C'^T$, $D' \geq 0$) will have a real symmetric solution as close as we wish to X_0 . (The precise definitions of various notions of stability are given in §3.) As a byproduct, we also obtain descriptions of stably disconjugate Hamiltonian systems of differential equations $Jx' = Hx$, where $H = H^T$ is a given constant matrix and $J = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}$. Often we will assume that the pair (A, D) is *controllable*, i.e., the rank of the $n \times n^2$ matrix $[D \ AD \ \cdots \ A^{n-1}D]$ is n . In the framework of complex matrices (in this case, A^T in (1.1) is replaced by the conjugate transpose A^*), the study of stability properties of Hermitian solutions was done in [RR2]. The approach we take in this paper to stability problems of solutions X_0 of (1.1) is based on reduction to corresponding stability problems of the graph subspace

$$\left\{ \begin{bmatrix} x \\ X_0 x \end{bmatrix} : x \in \mathbb{R}^n \right\}.$$

* Received by the editors January 10, 1990; accepted for publication (in revised form) January 17, 1991.

† Vrije Universiteit, Faculteit Wiskunde en Informatica, De Boelelaan 1081a, 1081 HV Amsterdam, the Netherlands.

‡ Department of Mathematics, College of William and Mary, Williamsburg, Virginia 23187-8795. The work of this author was partially supported by National Science Foundation grant DMS-8802836.

The graph subspace is M -invariant and J -neutral, where M is the Hamiltonian for (1.1). Various stability properties of such subspaces have been studied in [RR1].

The perturbation and stability properties of the solution of the algebraic Riccati equations have been studied in the literature in a somewhat different framework. Namely, assuming that the coefficients A , D , and C are C^r or analytic functions of a parameter (or several parameters), it can be shown under appropriate hypotheses that certain solutions of the algebraic Riccati equation are again C^r or analytic functions. See [D], [Ro2], and [RR4], where analytic dependence on parameters is assumed, and the main focus is on the stabilizing solution (in [RR4] the C^r -dependence is studied as well), and [Bu], where the C^1 -dependence is assumed (see also the last paragraph in § 3).

The paper is organized as follows. In the next section (which is of a preliminary character), we characterize (1.1) for which there is a real symmetric solution in terms of the canonical form of pair of matrices

$$\begin{bmatrix} A & D \\ C & -A^T \end{bmatrix}, \quad \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}.$$

The main results on stability and Lipschitz stability of real symmetric solutions of (1.1) are described in §§ 3 and 4, respectively. Some properties of stable solutions, particularly in relation to stability of nearby solutions of nearby equations and with isolatedness, are described in § 5. Finally, in the last section we apply these results to characterize stably disconjugate Hamiltonian systems.

We use the following notation and conventions throughout the paper. The spectrum (i.e., the set of eigenvalues, possibly complex) of an $n \times n$ real matrix T is denoted $\sigma(T)$. The algebraic multiplicity of an eigenvalue λ of T is defined to be the dimension of $\text{Ker}(T - \lambda I)^n$ (understood as a subspace in the complex space \mathbb{C}^n), and the geometric multiplicity is $\dim \text{Ker}(T - \lambda I)$. The partial multiplicities of T corresponding to λ are, by definition, the sizes of the Jordan blocks with eigenvalue λ in the Jordan form of T . Finally, $\text{Im } Z = \{Zx : x \in \mathbb{R}^n\} \subset \mathbb{R}^m$ denotes the column space of a real $m \times n$ matrix Z .

2. Existence of real symmetric solutions. A standard approach to the investigation of (1.1) is by introducing the $2n \times 2n$ real matrices

$$(2.1) \quad M = \begin{bmatrix} A & D \\ C & -A^T \end{bmatrix}; \quad J = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}.$$

We check that $J = -J^T$ and that $JM = -M^T J$. So $(M, J) \in L_{2n}(-1, -1)$, where we borrow the notation $L_p(-1, -1)$ from [RR1] to designate the set of all ordered pairs (B, H) of $p \times p$ real matrices B and H such that H is invertible, $H = -H^T$, and $HB = -B^T H$. (Observe that p must be even so that $L_p(-1, -1) \neq \emptyset$.) Such pairs (B, H) can be reduced to a canonical form (described, for example, in [DPWZ]) by simultaneous real congruence and similarity, as follows. We denote by $J_q(a)$ the lower triangular $q \times q$ Jordan block with eigenvalue a (which is assumed to be real), and by $J_q\left(\begin{smallmatrix} a & b \\ -b & a \end{smallmatrix}\right)$ the $2q \times 2q$ matrix

$$J_q\left(\begin{smallmatrix} a & b \\ -b & a \end{smallmatrix}\right) = \begin{bmatrix} Z & 0 & \cdots & 0 & 0 \\ I_2 & Z & \cdots & 0 & 0 \\ 0 & I_2 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & Z & 0 \\ 0 & 0 & \cdots & I_2 & Z \end{bmatrix}, \quad Z = \begin{bmatrix} a & b \\ -b & a \end{bmatrix};$$

here a and b are real numbers with $b \neq 0$, and I_2 denotes the 2×2 identity matrix.

THEOREM 2.1 ([DPWZ], [Th]). *Let $(B, H) \in L_p(-1, -1)$. Then there exists a real invertible matrix S such that the pair $(S^{-1}BS, S^THS) \in L_p(-1, -1)$ decomposes into a simultaneous direct sum*

$$(2.2) \quad S^{-1}BS = \bigoplus_{j=1}^p B_j, \quad S^THS = \bigoplus_{j=1}^p H_j,$$

where $(B_j, H_j) \in L_{p_j}(-1, -1)$ and each pair (B_j, H_j) is one of the following 5 types.

Type 1.

$$B_j = J_{2n_j}(0); \quad H_j = \kappa_j F_{2n_j},$$

where $\kappa_j = \pm 1$ and F_k is the $k \times k$ matrix given by

$$F_k = \begin{bmatrix} 0 & \cdots & 0 & 0 & 1 \\ 0 & & 0 & -1 & 0 \\ 0 & & 1 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ (-1)^{k-1} & \cdots & 0 & 0 & 0 \end{bmatrix}.$$

Type 2.

$$B_j = J_{2n_j+1}(0) \oplus (-J_{2n_j+1}(0))^T; \quad H_j = \begin{bmatrix} 0 & I_{2n_j+1} \\ -I_{2n_j+1} & 0 \end{bmatrix}.$$

Type 3.

$$B_j = J_{n_j}(a) \oplus (-J_{n_j}(a))^T; \quad H_j = \begin{bmatrix} 0 & I_{n_j} \\ -I_{n_j} & 0 \end{bmatrix},$$

where a is a positive real number.

Type 4.

$$B_j = J_{n_j} \begin{pmatrix} 0 & b \\ -b & 0 \end{pmatrix}; \quad H_j = \kappa_j \begin{bmatrix} 0 & \cdots & 0 & F_{2'}^{n_j} \\ 0 & \cdots & -F_{2'}^{n_j} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ (-1)^{n_j-1} & F_{2'}^{n_j} & \cdots & 0 & 0 \end{bmatrix},$$

where $b > 0$, $\kappa_j = \pm 1$, and $F_2 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$.

Type 5.

$$B_j = J_{n_j} \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \oplus \left(-J_{n_j} \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \right)^T; \quad H_j = \begin{bmatrix} 0 & I_{2n_j} \\ -I_{2n_j} & 0 \end{bmatrix}.$$

Moreover, the blocks (B_j, H_j) in (2.2) are uniquely determined by (B, H) up to a permutation.

The signs $\kappa_j = \pm 1$ that are attached to partial multiplicities corresponding to the pure imaginary ($\neq 0$) eigenvalues of B and to even partial multiplicities corresponding to the zero eigenvalue of B are also uniquely determined (up to permutation of signs attached to equal partial multiplicities corresponding to the same eigenvalue of B) and form the *sign characteristic* of (B, H) .

We now return to the Riccati equation (1.1) and the matrices (M, J) given by (2.1). The basic existence result is given by the following theorem.

THEOREM 2.2. *Assume that $C = C^T$, $D = D^T \geq 0$, and the pair (A, D) is controllable, i.e., the rank of the $n \times n^2$ matrix $[D \ AD \ \cdots \ A^{n-1}D]$ is n . Then the following statements are equivalent:*

- (i) *there exists a real symmetric solution X of (1.1);*

- (ii) all pure imaginary and zero eigenvalues of M (if any) have only even partial multiplicities and the signs κ_i in the sign characteristic of (M, J) corresponding to the (nonzero) pure imaginary eigenvalues are all -1 , while the sign κ_j attached to the even partial multiplicity m corresponding to the zero eigenvalue is $(-1)^{m/2}$;
- (iii) all pure imaginary and zero eigenvalues of M (if any) have only even partial multiplicities.

The equivalence (i) \Leftrightarrow (iii) in Theorem 2.2 is well known (see [Cu], [LR], and [S]). The implication (ii) \Rightarrow (iii) is evident, and, finally, the implication (i) \Rightarrow (ii) follows from Lemma 2.3 below. The proof of this lemma uses basically known ideas. Close results are found in [Ro1] and [GLR4], but in this form the result of Lemma 2.3 appears to be new.

LEMMA 2.3. Let A , D , and C be real $n \times n$ matrices with $C = C^T$, $D = D^T$, D positive semidefinite and (A, D) controllable. If there is a real matrix $X = X^T$ such that

$$XDX + XA + A^T X - C = 0,$$

then every pure imaginary or zero eigenvalue of

$$M := \begin{pmatrix} A & D \\ C & -A^T \end{pmatrix}$$

has even partial multiplicities, and the signs in the sign characteristic of $(M, \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix})$ corresponding to pure imaginary nonzero eigenvalues are all -1 , while the sign κ_j attached to the partial multiplicity m corresponding to the zero eigenvalue is $(-1)^{m/2}$.

Proof. Let

$$J = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}.$$

We verify that

$$(2.3) \quad \begin{pmatrix} I & 0 \\ -X & I \end{pmatrix} M \begin{pmatrix} I & 0 \\ X & I \end{pmatrix} = \begin{pmatrix} A + DX & D \\ 0 & -(A + DX)^T \end{pmatrix}$$

and

$$(2.4) \quad \begin{pmatrix} I & X \\ 0 & I \end{pmatrix} J \begin{pmatrix} I & 0 \\ X & I \end{pmatrix} = J.$$

Furthermore, let $Z = S^{-1}(A + DX)S$ be the real Jordan form of $A + DX$ (here S is some real invertible matrix). We verify that

$$(2.5) \quad \begin{pmatrix} S^{-1} & 0 \\ 0 & S^T \end{pmatrix} \begin{pmatrix} A + DX & D \\ 0 & -(A + DX)^T \end{pmatrix} \begin{pmatrix} S & 0 \\ 0 & S^{-1T} \end{pmatrix} = \begin{pmatrix} Z & D_0 \\ 0 & -Z^T \end{pmatrix},$$

where $D_0 = S^{-1}DS^{-1T}$ is symmetric and positive semidefinite; and

$$(2.6) \quad \begin{pmatrix} S^T & 0 \\ 0 & S^{-1} \end{pmatrix} J \begin{pmatrix} S & 0 \\ 0 & S^{-1T} \end{pmatrix} = J.$$

It is easily seen that the pair (Z, D_0) is controllable as well. Because of formulas (2.3)–(2.5), we can consider the pair

$$\left(\begin{bmatrix} Z & D_0 \\ 0 & -Z^T \end{bmatrix}, J \right)$$

in place of the original pair (M, J) . Let ib ($b \geq 0$) be a pure imaginary (or zero) eigenvalue of Z . Without loss of generality we can write $Z = Z_1 \oplus Z_2$, where $\sigma(Z_1) = \{\pm ib\}$; $\sigma(Z_2) \cap \{\pm ib\} = \emptyset$. Partition D_0 accordingly as follows:

$$D_0 = \begin{pmatrix} D_{01} & D_{02} \\ D_{02}^T & D_{03} \end{pmatrix}.$$

Apply the following similarity transformation to $\begin{bmatrix} Z & D_0 \\ 0 & -Z^T \end{bmatrix}$:

$$\begin{pmatrix} I & 0 & 0 & -U \\ 0 & I & -U^T & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{pmatrix} \begin{pmatrix} Z_1 & 0 & D_{01} & D_{02} \\ 0 & Z_2 & D_{02}^T & D_{03} \\ 0 & 0 & -Z_1^T & 0 \\ 0 & 0 & 0 & -Z_2^T \end{pmatrix} \begin{pmatrix} I & 0 & 0 & U \\ 0 & I & U^T & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{pmatrix} \\ = \begin{pmatrix} Z_1 & 0 & D_{01} & \tilde{D}_{02} \\ 0 & Z_2 & \tilde{D}_{02}^T & D_{03} \\ 0 & 0 & -Z_1^T & 0 \\ 0 & 0 & 0 & -Z_2^T \end{pmatrix},$$

where the real matrix U is chosen so that $\tilde{D}_{02} = 0$ (it is easy to see that such choice is possible because $\sigma(Z_1) \cap \sigma(Z_2) = \emptyset$). We verify that

$$\begin{pmatrix} I & 0 \\ V & I \end{pmatrix} J \begin{pmatrix} I & V \\ 0 & I \end{pmatrix} = J,$$

where

$$V = V^T = \begin{pmatrix} 0 & U \\ U^T & 0 \end{pmatrix},$$

and that the pair (Z_1, D_{01}) is controllable.

We have reduced the proof of the lemma to the following situation: either

$$(2.7) \quad M = \begin{bmatrix} \bigoplus_{j=1}^p J_{n_j} \begin{pmatrix} 0 & b \\ -b & 0 \end{pmatrix} & D \\ 0 & \bigoplus_{j=1}^p \left(-J_{n_j} \begin{pmatrix} 0 & b \\ -b & 0 \end{pmatrix} \right)^T \end{bmatrix}$$

(where $b > 0$), or

$$(2.8) \quad M = \begin{bmatrix} \bigoplus_{j=1}^p J_{n_j}(0) & D \\ 0 & \bigoplus_{j=1}^p (-J_{n_j}(0))^T \end{bmatrix}.$$

First, consider the case where M is given by (2.8). Using Lemma 8.4 of [GLR1] or Lemma 3.4 of [GLR2] we verify that the partial multiplicities of M are $2n_1, \dots, 2n_p$. Observe that, for a given real $n \times n$ matrix K , the set \mathcal{V} of all real symmetric positive semidefinite $n \times n$ matrices V such that (K, V) is controllable, is connected (this follows, for example, from the easily verified fact that (K, V) is controllable if and only if the

matrix

$$[I \ K \cdots K^{n-1}] \begin{bmatrix} V & 0 & \cdots & 0 \\ 0 & V & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & V \end{bmatrix} \begin{bmatrix} I \\ K^T \\ \vdots \\ (K^T)^{n-1} \end{bmatrix}$$

is positive definite). Let $K = \bigoplus_{j=1}^p J_{n_j}(o)$. Because of the connectivity of \mathcal{V} , Theorem II.1.1 of [GLR3] implies that the sign characteristic of

$$\left(\begin{bmatrix} K & V \\ 0 & -K^T \end{bmatrix}, \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix} \right)$$

is independent of $V \in \mathcal{V}$ (Theorem II.1.1 of [GLR3] is stated for the framework of complex matrices; however, Proposition 3.3 of [RR1] makes it possible to apply this theorem for the real case as well). Thus, the sign characteristic of (M, J) is the same as that of

$$(2.9) \quad \left(\begin{bmatrix} K & V_0 \\ 0 & -K^T \end{bmatrix}, J \right),$$

where $V_0 = e_1 e_1^T + e_{n_1+1} e_{n_1+1}^T + \cdots + e_{n_1+\cdots+n_{p-1}+1} e_{n_1+\cdots+n_{p-1}+1}^T$. (We denote by e_k the column vector with 1 in the k th position and zeros elsewhere; the dimension of e_k is clear from the context.)

By inspection, we verify that the signs κ_j in the sign characteristic of the pair (2.9) are $(-1)^{n_1}, \dots, (-1)^{n_p}$. (This part of the proof follows the arguments used in the proof of Theorem 4 in [Ro1].)

Consider now the case where M is given by (2.7). Let

$$K = \bigoplus_{j=1}^p J_{n_j} \begin{bmatrix} 0 & b \\ -b & 0 \end{bmatrix} \quad (b > 0).$$

The existence of a Hermitian solution (namely, the zero solution) of the Riccati equation $XXD + XK + K^T X = 0$ implies (in view of the results of [Cu], [LR], and [S]; see also Theorem II.4.3 in [GLR4]) that all the partial multiplicities of M are even. Actually, the proof of Theorem II.4.3 in [GLR4] shows that the partial multiplicities of M corresponding to the eigenvalue $\pm ib$ are $2n_1, \dots, 2n_p$. Theorem 4 in [Ro1] and Proposition I.6.8 in [GLR4] combined tell us that the sign characteristic of the pair

$$\left(i \begin{bmatrix} K & D \\ 0 & -K^T \end{bmatrix}, i \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix} \right)$$

(as defined in, e.g., [GLR4]), for complex matrices selfadjoint in a complex indefinite scalar product, consists entirely of -1 's. Proposition 3.3 in [RR1] implies that all the signs in the sign characteristic of

$$\left(i \begin{bmatrix} K & D \\ 0 & -K^T \end{bmatrix}, i \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix} \right)$$

are -1 's. \square

The conditions (i)–(iii) in Theorem 2.2 are equivalent to the existence of an M -invariant J -Lagrangian subspace (see, e.g., [F], where several other equivalent conditions are stated under the somewhat milder assumption of sign-controllability of (A, D)). In general, given a real matrix N such that $JN = -N^T J$, criteria for existence of an N -invariant J -Lagrangian subspace are given in [CD] (the context in [CD] is

different from ours) and in [RR1]. The criterion given in [RR1] states that there is an N -invariant J -Lagrangian subspace if and only if for every eigenvalue ib ($b > 0$) of N the number of odd partial multiplicities corresponding to ib is even, and the sum of the signs κ_j attached to these odd partial multiplicities is zero. However, the situation in Theorem 2.2 is special (because of the hypotheses that $D \geq 0$ and (A, D) is controllable, which need not be satisfied for a general matrix

$$N = \begin{bmatrix} A & D \\ C & -A^T \end{bmatrix}$$

such that $JN = -N^T J$), and therefore the criterion given in Theorem 2.2(ii) for existence of an M -invariant J -Lagrangian subspace also takes a special form.

3. Stable symmetric solutions. We introduce the concept of stable solutions of the algebraic Riccati equation (1.1). It is convenient to first introduce some notation. Let \mathcal{C}_n be the set of all triples of real $n \times n$ matrices (A, C, D) such that $C = C^T$, $D = D^T$; D is positive semidefinite, and the pair (A, D) is controllable. For $(A, C, D) \in \mathcal{C}_n$ let $\chi(A, C, D)$ be the set of real symmetric solutions (possibly an empty set) of the algebraic Riccati equation

$$(3.1) \quad XDX + XA + A^T X - C = 0.$$

A solution $X \in \chi(A, C, D)$ will be called *conditionally stable* if for every $\varepsilon > 0$ there is $\delta > 0$ such that every triple $(A', C', D') \in \mathcal{C}_n$ for which $\chi(A', C', D') \neq \emptyset$ and for which

$$\|A - A'\| + \|C - C'\| + \|D - D'\| < \delta$$

has a solution $X' \in \chi(A', C', D')$ with $\|X' - X\| < \varepsilon$. If the condition $\chi(A', C', D') \neq \emptyset$ is omitted from the above definition, we obtain the definition of an *unconditionally stable* solution X .

A standard approach (also adopted here) to study the solutions X of (3.1) is by way of the graph subspaces

$$(3.2) \quad \mathcal{N}(X) = \text{Im} \begin{bmatrix} I \\ X \end{bmatrix} \subset \mathbb{R}^{2n}.$$

It is easy to see that $\mathcal{N}(X)$ is M -invariant and J -Lagrangian for every real symmetric solution X of (3.1); here

$$(3.3) \quad M = \begin{bmatrix} A & D \\ C & -A^T \end{bmatrix}, \quad J = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}.$$

Conversely (see [Cu], [LR], and [S], and also Theorem 2.2), if (A, D) is controllable, then every M -invariant J -Lagrangian subspace is of the form $\text{Im} \begin{bmatrix} I \\ X \end{bmatrix}$ for some real symmetric solution of (3.1).

Next, we need the notion of (un)conditionally stable invariant Lagrangian subspaces, introduced and studied in [RR1]. For subspaces $\mathcal{M}, \mathcal{N} \subset \mathbb{R}^p$, we let $\Theta(\mathcal{M}, \mathcal{N})$ be the *gap metric* defined by

$$\Theta(\mathcal{M}, \mathcal{N}) = \|P_{\mathcal{M}} - P_{\mathcal{N}}\|,$$

where $P_{\mathcal{M}}(P_{\mathcal{N}})$ is the orthogonal projection on $\mathcal{M}(\mathcal{N})$, with respect to the standard inner product

$$\langle (x_1, \dots, x_p)^T, (y_1, \dots, y_p)^T \rangle = \sum_{j=1}^p x_j y_j$$

in \mathbb{R}^p . Let $(B, H) \in L_p(-1, -1)$, and let \mathcal{N} be a B -invariant H -Lagrangian subspace in \mathbb{R}^p . We say that \mathcal{N} is *conditionally stable* if for every $\varepsilon > 0$ there is $\delta > 0$ such that every pair $(B', H') \in L_p(-1, -1)$ with $\|B - B'\| + \|H - H'\| < \delta$ has a B' -invariant H' -Lagrangian subspace \mathcal{N}' with $\Theta(\mathcal{N}, \mathcal{N}') < \varepsilon$ *provided* the set of B' -invariant H' -Lagrangian subspaces is not empty. If this proviso is removed, we obtain the definition of an *unconditionally stable* B -invariant H -Lagrangian subspace \mathcal{N} .

In the following, for a given $m \times m$ real matrix \mathcal{Z} , we denote by $R(\mathcal{Z}, \lambda)$ the root subspace of \mathcal{Z} corresponding to its real eigenvalue λ , and by $R(\mathcal{Z}, a \pm ib)$ the real root subspace of \mathcal{Z} corresponding to the pair of complex conjugate nonreal eigenvalue $a \pm ib$ as follows:

$$R(\mathcal{Z}, \lambda) = \text{Ker}(\mathcal{Z} - \lambda I)^m,$$

$$R(\mathcal{Z}, a \pm ib) = \text{Ker}(\mathcal{Z}^2 - 2a\mathcal{Z} + (a^2 + b^2)I)^m.$$

LEMMA 3.1. *Let $(A, C, D) \in \mathcal{C}_n$, and let*

$$M = \begin{bmatrix} A & D \\ C & -A^T \end{bmatrix}, \quad J = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}.$$

Then $X \in \chi(A, C, D)$ is an (un)conditionally stable solution of (3.1) if and only if $\mathcal{N}(X)$ is (un)conditionally stable as an M -invariant J -Lagrangian subspace.

Proof. Clearly, if $\mathcal{N}(X)$ is (un)conditionally stable, then X is an (un)conditionally stable solution of (3.1).

Conversely, assume $\mathcal{N}(X)$ is not (un)conditionally stable. We must prove that X is not (un)conditionally stable. Let X_0 be a real symmetric solution of (3.1) such that $\sigma(A + DX_0)$ lies in the closed left half-plane. (The existence of such X_0 is ensured because $\chi(A, C, D)$ is nonempty; see, e.g., [GLR4]. Note also that X_0 is unique, by Theorem 2.1 and Theorem 3.2 in [RR1].) Furthermore, let

$$\tilde{M} = \begin{bmatrix} A + DX_0 & D \\ 0 & -(A + DX_0)^T \end{bmatrix}.$$

Applying similarity (2.3) (with X replaced by X_0) it is easy to see that the \tilde{M} -invariant J -Lagrangian subspace

$$\mathcal{N}_0 = \text{Im} \begin{bmatrix} I \\ X - X_0 \end{bmatrix}$$

is not (un)conditionally stable.

At this point, it is convenient to separately consider the two concepts of stability. So first assume that \mathcal{N}_0 is not conditionally stable. By Theorem 3.4 of [RR1], we conclude that at least one of three following conditions holds:

- (a) $(0) \neq \mathcal{N}_0 \cap R(\tilde{M}, \lambda_0) \neq R(\tilde{M}, \lambda_0)$ for some real nonzero eigenvalue λ_0 of \tilde{M} such that $\dim \text{Ker}(\tilde{M} - \lambda_0 I) > 1$;
- (b) $\mathcal{N}_0 \cap R(\tilde{M}, \lambda_0)$ is odd-dimensional for some real nonzero eigenvalue λ_0 of \tilde{M} such that $\dim \text{Ker}(\tilde{M} - \lambda_0 I) = 1$ and $R(\tilde{M}, \lambda_0)$ is even-dimensional;
- (c) $0 \neq \mathcal{N}_0 \cap R(\tilde{M}, a_0 \pm ib_0) \neq R(\tilde{M}, a_0 \pm ib_0)$ for some nonreal nonpure imaginary eigenvalues $a_0 \pm ib_0$ ($a_0 \neq 0, b_0 > 0$) of \tilde{M} with geometric multiplicity at least 2.

Observe that since $(\tilde{M}, J) \in L(-1, -1)$ (see also Theorem 3.4 in [RR1]), if any of the conditions (a), (b), or (c) is satisfied for λ_0 or $a_0 \pm ib_0$ as appropriate, then the same condition is satisfied for $-\lambda_0$ or $-a_0 \pm ib_0$. So we can assume $\lambda_0 < 0$ or $a_0 < 0$, as appropriate.

Because of our choice of X_0 , the subspace $R(\tilde{M}, \lambda_0)$, or $R(\tilde{M}, a_0 \pm ib_0)$, is contained in $\text{Im} \begin{bmatrix} I \\ 0 \end{bmatrix}$. In view of the description of stable invariant subspaces for real matrices (see [BGK], [GLR2], or [GLR5]), there is a sequence of real matrices $\{A_m\}_{m=1}^\infty$ such that $A_m \rightarrow A + DX_0$ as $m \rightarrow \infty$ and for every A_m -invariant subspace $\mathcal{N}_m (\subset \mathbb{R}^n)$

$$(3.4) \quad \Theta \left(\begin{bmatrix} \mathcal{N}_m \\ 0 \end{bmatrix}, \mathcal{N}_0 \cap \mathcal{F} \right) \geq \varepsilon_0,$$

where ε_0 is independent of m , and \mathcal{F} is the sum of all root subspaces of \tilde{M} corresponding to the eigenvalues in the open left half-plane. Clearly, $(A_m, 0, D) \in \mathcal{C}_n$ for m large enough and $\chi(A_m, 0, D) \neq \emptyset$. Let

$$(3.5) \quad \tilde{M}_m = \begin{bmatrix} A_m & D \\ 0 & -A_m^T \end{bmatrix}.$$

We claim that for any \tilde{M}_m -invariant J -Lagrangian subspace $\tilde{\mathcal{N}}_m$ the inequality

$$(3.6) \quad \Theta(\tilde{\mathcal{N}}_m, \mathcal{N}_0) \geq \varepsilon_1$$

holds (here, and in the following, we denote by $\varepsilon_1, \varepsilon_2, \dots$ positive constants independent of m). Indeed, if (3.6) were false, then we would have

$$(3.7) \quad \Theta(\tilde{\mathcal{N}}_m \cap \mathcal{F}_m, \mathcal{N}_0 \cap \mathcal{F}) \rightarrow 0 \quad \text{as } m \rightarrow \infty,$$

where \mathcal{F}_m is the \tilde{M}_m -invariant subspace equal to the sum of all root subspaces of \tilde{M}_m corresponding to eigenvalues that converge (as $m \rightarrow \infty$) to the open left half-plane eigenvalues of \tilde{M} . The form (3.5) of \tilde{M} easily implies that

$$\mathcal{F}_m \subset \text{Im} \begin{bmatrix} I \\ 0 \end{bmatrix},$$

(i.e., we can consider $\tilde{\mathcal{N}}_m \cap \mathcal{F}_m$ as an A_m -invariant subspace), and (3.7) clearly contradicts (3.4).

We conclude from (3.6) that every solution $X_m = X_m^T$ of the equation

$$(3.8) \quad YDY + YA_m + A_m^T Y = 0$$

satisfies

$$(3.9) \quad \|X_m - (X - X_0)\| \geq \varepsilon_2.$$

Consider now the equation

$$(3.10) \quad \begin{aligned} & ZDZ + Z(A_m - DX_0) + (A_m - DX_0)^T Z \\ & + X_0 DX_0 - X_0 A_m - A_m^T X_0 = 0. \end{aligned}$$

We easily verify that Y is a symmetric solution of (3.8) if and only if $Y + X_0$ is a symmetric solution of (3.10). Also, $A_m - DX_0 \rightarrow A$ and

$$X_0 DX_0 - X_0 A_m - A_m^T X_0 \rightarrow -C$$

as $m \rightarrow \infty$. Now (3.9) implies that X is not conditionally stable.

Consider now the unconditional stability. So let \mathcal{N}_0 not be unconditionally stable. If $\sigma(\tilde{M}) \cap i\mathbb{R} = \emptyset$, then, in view of Theorem 3.4 of [RR1], we can argue as above to prove that X is not unconditionally stable. It remains to be proved (by the same Theorem 3.4 of [RR1]) that if \tilde{M} has pure imaginary or zero eigenvalues, then there is no unconditionally stable solution X .

It is convenient to reformulate the statement that we will prove in the following form: Given $(A, 0, D) \in \mathcal{C}_n$ with $\operatorname{Re} \sigma(A) \leq 0$, and such that A has pure imaginary or zero eigenvalues. For every $\varepsilon > 0$ find $(A', C', D') \in \mathcal{C}_n$ such that $\|A' - A\| + \|C'\| + \|D' - D\| < \varepsilon$ and $\chi(A', C', D') = \emptyset$. Applying the reductions as in the proof of Lemma 2.3, we can assume without loss of generality that the matrix

$$M := \begin{bmatrix} A & D \\ 0 & A^T \end{bmatrix}$$

is given by (2.7) or (2.8).

First, consider the case where M is given by (2.8). Because of the controllability condition, the top left corner entry d of D is positive. Let $C'(\varepsilon)$ be the $n \times n$ matrix with $-\varepsilon$ in the top left corner and zeros elsewhere, and let

$$M(\varepsilon) = \begin{bmatrix} A & d \\ C'(\varepsilon) & -A^T \end{bmatrix}.$$

Clearly, $(A, C'(\varepsilon), D) \in \mathcal{C}_n$. It is not difficult to check that for $\varepsilon > 0$ the matrix $M(\varepsilon)$ has $2n - 2$ zero eigenvalues (counting multiplicities) and the simple eigenvalues $\pm i\sqrt{\varepsilon d}$. Indeed, by an appropriate similarity, $M(\varepsilon)$ can be reduced to the form

$$\begin{bmatrix} 0 & d & 0 & * \\ -\varepsilon & 0 & & \\ 0 & J_{n_1-1}(0) \oplus \bigoplus_{j=2}^p J_{n_j}(0) & & * \\ 0 & 0 & -J_{n_1-1}(0)^T \oplus \bigoplus_{j=2}^p J_{n_j}^T(0) \end{bmatrix}.$$

By Theorem 2.2, $\chi(A, C'(\varepsilon), D) = \emptyset$ for $\varepsilon > 0$.

Finally, assume M is given by (2.7). Let us first verify that for any positive definite real symmetric matrix

$$\begin{bmatrix} d_1 & d_2 \\ d_2 & d_3 \end{bmatrix}$$

and for any $b > 0$ the 4×4 matrix

$$Q(\varepsilon) = \begin{bmatrix} 0 & b & d_1 & d_2 \\ -b & 0 & d_2 & d_3 \\ -\varepsilon & 0 & 0 & b \\ 0 & -\varepsilon & -b & 0 \end{bmatrix}$$

has four distinct pure imaginary (nonzero) eigenvalues different from $\pm ib$ when $\varepsilon > 0$ is sufficiently close to zero. Indeed, we can assume without loss of generality that $b = 1$; then the characteristic polynomial of $Q(\varepsilon)$ is

$$\lambda^4 + 2\lambda^2(1 + \varepsilon q) + 1 - 2\varepsilon q + \varepsilon^2 r = (\lambda^2 + (1 + \varepsilon q))^2 + \varepsilon^2(r - q^2) - 4\varepsilon q,$$

where

$$q = \frac{1}{2}(d_1 + d_3); \quad r = d_1 d_3 - d_2^2.$$

Since $q > 0$,

$$\varepsilon^2(r - q^2) - 4\varepsilon q < 0 \quad (\text{for } \varepsilon > 0)$$

and

$$(1 + \varepsilon q)^2 > \varepsilon^2(r - q^2) - 4\varepsilon q \quad (\text{for all real } \varepsilon),$$

the conclusion concerning eigenvalues of $Q(\varepsilon)$ follows. We now apply a perturbation similar to that applied in the case M , given by (2.8). Let $C'(\varepsilon)$ be the $n \times n$ matrix with $-\varepsilon$ in the $(1, 1)$ and $(2, 2)$ positions, and zeros elsewhere (here $\varepsilon > 0$). Again, $(A, C'(\varepsilon), D) \in \mathcal{C}_n$, and the matrix

$$M(\varepsilon) = \begin{bmatrix} A & D \\ C'(\varepsilon) & -A^T \end{bmatrix}$$

can be reduced by a suitable similarity to the form

$$\begin{bmatrix} \begin{matrix} 0 & b & d_1 & d_2 \\ -b & 0 & d_2 & d_3 \\ -\varepsilon & 0 & 0 & b \\ 0 & -\varepsilon & -b & 0 \end{matrix} & \begin{matrix} 0 \\ * \\ * \\ 0 \end{matrix} \\ \begin{matrix} 0 \\ 0 \end{matrix} & \begin{matrix} J_{n_1-1} \begin{bmatrix} 0 & b \\ -b & 0 \end{bmatrix} \oplus \bigoplus_{j=2}^p J_{n_j} \begin{bmatrix} 0 & b \\ -b & 0 \end{bmatrix} \\ 0 & -J_{n_1-1} \begin{bmatrix} 0 & b \\ -b & 0 \end{bmatrix}^T \oplus \bigoplus_{j=2}^p -J_{n_j} \begin{bmatrix} 0 & b \\ -b & 0 \end{bmatrix}^T \end{matrix} \end{bmatrix}.$$

Here

$$\begin{bmatrix} d_1 & d_2 \\ d_2 & d_3 \end{bmatrix}$$

is the top left 2×2 corner of D , which is positive definite in view of the controllability condition. Now $M(\varepsilon)$ has simple pure imaginary eigenvalues (when $\varepsilon > 0$ is close to zero), so $\chi(A, C'(\varepsilon), D) = \emptyset$ by Theorem 2.2, and we are done. \square

With the results of Lemma 3.1, the proof of the following descriptions of all stable symmetric solutions of the algebraic Riccati equation is done simply by appealing to Theorem 3.4 of [RR1].

THEOREM 3.2. *Let*

$$(3.11) \quad XDX + XA + A^T X - C = 0$$

be an algebraic Riccati equation with real $n \times n$ matrices A , $C = C^T$, and $D = D^T$ such that D is positive semidefinite and (A, D) is a controllable pair, and let

$$M = \begin{bmatrix} A & D \\ C & -A^T \end{bmatrix}.$$

A solution $X = X^T$ (X is a real matrix) of (3.11) is conditionally stable if and only if the following conditions are satisfied:

- (i) $\text{Im} \left[\begin{smallmatrix} I \\ X \end{smallmatrix} \right] \cap R(M, \lambda)$ is either zero or $R(M, \lambda)$ whenever $\lambda \neq 0$ is a real eigenvalue of M with geometric multiplicity greater than 1;
- (ii) $\text{Im} \left[\begin{smallmatrix} I \\ X \end{smallmatrix} \right] \cap R(M, \lambda)$ is an even-dimensional subspace whenever $\lambda \neq 0$ is a real eigenvalue of M of geometric multiplicity 1 and even algebraic multiplicity;
- (iii) $\text{Im} \left[\begin{smallmatrix} I \\ X \end{smallmatrix} \right] \cap R(M, a \pm ib)$ is either zero or $R(M, a \pm ib)$ whenever $a \pm ib$ is a conjugate pair of nonreal nonpure imaginary eigenvalues of M with geometric multiplicity at least 2.

THEOREM 3.3. *In the notation and under the hypotheses of Theorem 3.2, there exists an unconditionally stable real symmetric solution X if and only if M has no pure imaginary*

or zero eigenvalues. If this condition is satisfied, then X is unconditionally stable if and only if it is conditionally stable.

The first part of this theorem also follows from Corollary 3 in [F2].

Of special interest are the maximal solution X_+ and the minimal solution X_- (i.e., X_+ and X_- are real symmetric solutions of (3.11) such that $X_+ - X$ and $X - X_-$ are positive semidefinite for every real symmetric solution X). If the set of real symmetric solutions is nonempty, then X_+ and X_- always exist and are characterized by the properties that $\sigma(A + DX_+)$ lies in the closed left half-plane, and $\sigma(A + DX_-)$ lies in the closed right half-plane (see [K], [KS], [LR], and [S]).

COROLLARY 3.4. *The maximal and minimal solutions X_+ and X_- are conditionally stable. If M has no pure imaginary or zero eigenvalues, then X_+ and X_- are unconditionally stable.*

Actually, in the case where M has no pure imaginary or zero eigenvalues, the solutions X_+ and X_- are unconditionally Lipschitz stable as well (see Theorem 4.2 in the next section).

It is instructive to compare Theorems 3.2 and 3.3 with the results on structural stability of solution of the algebraic Riccati equations proved in [Bu]. A (real symmetric) solution X of (1.1) is called *structurally stable* if the following holds: Given any continuously differentiable functions $A(t)$, $C(t) = C(t)^*$, and $D(t) = D(t)^*$ such that $A(t_0) = A$, $C(t_0) = C$, and $D(t_0) = D$, for $|t - t_0|$ sufficiently small there exists a unique (in a sufficiently small neighborhood of X) solution $X(t)$ of the equation $YD(t)Y + YA(t) + A(t)^T Y - C(t) = 0$, and, moreover, the number of eigenvalues (counted with multiplicities) of $A(t) + D(t)X(t)$ in each of the three regions—open left half-plane, imaginary axis, open right half-plane—coincides with the number of eigenvalues of $A + DX$ in the corresponding region. The crucial difference with our definition of unconditionally stable solutions is that in the case of a structurally stable solution X , the perturbed equation must have a *unique* solution in a neighborhood of X . As proved in [Bu], a solution X is structurally stable if and only if $\sigma(A + DX) \cap \sigma(-A^T - XD) = \emptyset$, i.e., if and only if X is a Lipschitz stable solution (see Theorem 4.2 in the next section). In this connection we point out that an invariant subspace \mathcal{M} of a complex matrix A is Lipschitz stable if and only if every nearby matrix has a unique invariant subspace in a suitable neighborhood of \mathcal{M} (see [RR3] and Theorem 15.5.1 in [GLR5]); no symmetries are assumed here.

4. Lipschitz stability. Consider (3.1) with the usual (in this paper) assumption that $(A, C, D) \in \mathcal{C}_n$. A real symmetric solution X of (3.1) will be called *conditionally Lipschitz stable* if there exist positive constants K, ε such that for every triple $(A', C', D') \in \mathcal{C}_n$ satisfying $\|A - A'\| + \|C - C'\| + \|D - D'\| < \varepsilon$, and with $\chi(A', C', D') \neq \emptyset$ there is a real symmetric X' with the properties that $X'D'X' + X'A' + A'^T X' - C' = 0$ and $\|X - X'\| \leq K(\|A - A'\| + \|C - C'\| + \|D - D'\|)$. If the statement “and with $\chi(A', C', D') \neq \emptyset$ ” is omitted from this definition, we obtain the notion of *unconditionally Lipschitz stable* X .

THEOREM 4.1. *Assume $(A, C, D) \in \mathcal{C}_n$. Then the following statements are equivalent:*

- (i) *there exists a conditionally Lipschitz stable real symmetric solution of (3.1);*
- (ii) *there exists an unconditionally Lipschitz stable real symmetric solution of (3.1);*
- (iii) *the matrix*

$$M := \begin{bmatrix} A & D \\ C & -A^T \end{bmatrix}$$

has no pure imaginary or zero eigenvalues.

Proof. Clearly (ii) \Rightarrow (i). Assume that (iii) holds. By Theorem 3.5 of [RR1], there exists an unconditionally Lipschitz stable M -invariant J -Lagrangian subspace \mathcal{N} (as before, $J = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}$). The subspace \mathcal{N} must be of the form $\text{Im} \begin{bmatrix} I \\ X \end{bmatrix}$, where X is a real symmetric solution of (3.1) (cf. the proof of Theorem 2.2). Then X is unconditionally Lipschitz stable, thereby proving (ii).

It remains to be proved that (i) implies (iii). Let X be a conditionally Lipschitz stable solution of (3.1), put

$$\tilde{M} = \begin{bmatrix} I & 0 \\ -X & I \end{bmatrix} \begin{bmatrix} A & D \\ C & -A^T \end{bmatrix} \begin{bmatrix} I & 0 \\ X & I \end{bmatrix} = \begin{bmatrix} A+DX & D \\ 0 & -(A+DX)^T \end{bmatrix} = \begin{bmatrix} \tilde{A} & D \\ 0 & -\tilde{A}^T \end{bmatrix}.$$

Then 0 is a conditionally Lipschitz stable solution of $XDX + X\tilde{A} + \tilde{A}^T X = 0$.

So we can restrict our attention to the case $C = 0$ and the zero solution of

$$(4.1) \quad XDX + XA + A^T X = 0.$$

We assume that A has pure imaginary or zero eigenvalues, and it will be proved that the zero solution of (4.1) is not conditionally Lipschitz stable.

Let $E_1 \geq 0$ and $E_2 < 0$ be $n \times n$ real symmetric matrices such that the $2n \times 2n$ matrix

$$\begin{bmatrix} 0 & E_1 \\ -E_2 & 0 \end{bmatrix}$$

has $2n$ distinct eigenvalues. Consider the Riccati equation

$$(4.2) \quad X(D + \alpha E_1)X + XA + A^T X + \alpha E_2 = 0,$$

for every $\alpha \geq 0$. Put $R(X) = -X(D + \alpha E_1)X - XA - A^T X - \alpha E_2$. Then $R(0) = -\alpha E_2 \geq 0$. By Theorem 2.1 in [RV] it follows that (4.2) has Hermitian solutions, and therefore also has real symmetric solutions. Note in this connection that the controllability of (A, D) implies the controllability of $(A, D + \alpha E_1)$. Let

$$M(\alpha) = \begin{bmatrix} A & D + \alpha E_1 \\ -\alpha E_2 & -A^T \end{bmatrix}$$

be the matrix corresponding to (4.2). Since for large α the matrix $M(\alpha)$ has $2n$ distinct eigenvalues, by a general result on the Jordan structure of a matrix analytically depending on a parameter (see [B], [GLR2]), for $\alpha_0 > 0$ sufficiently small the matrix $M(\alpha)$, $0 < \alpha \leq \alpha_0$, also has $2n$ distinct eigenvalues.

In particular, the number of $M(\alpha)$ -invariant J -Lagrangian subspaces is finite and is independent of α , for $0 < \alpha \leq \alpha_0$. So the number of real symmetric solutions $X(\alpha)$ of (4.2) is also finite and independent of α ($0 < \alpha \leq \alpha_0$). Moreover, the Puiseux expansions of eigenvalues, eigenvectors, and generalized eigenvectors of a matrix depending analytically on a parameter (see [B] and [GLR2]), together with the formula

$$\mathcal{N}(\alpha) = \text{Im} \begin{bmatrix} I \\ X(\alpha) \end{bmatrix},$$

where $\mathcal{N}(\alpha)$ is any $M(\alpha)$ -invariant J -Lagrangian subspace, show that $X(\alpha)$ is given by a fractional power series

$$(4.3) \quad X(\alpha) = \sum_{j=-k}^{\infty} \alpha^{j/p} X_j,$$

for some choice of the p th root $\alpha^{1/p}$ (here k is a nonnegative integer, and X_j are $n \times n$ matrices that may depend on the choice of $\alpha^{1/p}$). Arguing by contradiction, assume

that 0 is a conditionally Lipschitz stable real symmetric solution of (4.1). Then there is $K > 0$ such that for every $\alpha \in (0, \alpha_0]$ there is a solution $\tilde{X}(\alpha)$ of (4.2) satisfying

$$(4.4) \quad \|\tilde{X} - \tilde{X}(\alpha)\| \leq K\alpha.$$

As the number of branches in (4.3) is finite, there will be a sequence $\{\alpha_j\}_{j=1}^\infty$ of positive numbers such that $\lim_{j \rightarrow \infty} \alpha_j = 0$ and $\tilde{X}(\alpha_j)$ belongs to the same branch in (4.3). For this branch (4.4) implies that $k=0$, $X_0 = \tilde{X}$, and $X_1 = \cdots = X_{p-1} = 0$ in (4.3). In particular, $X(\alpha)$ is differentiable at $\alpha=0$ along this branch, and $X'(0) = X_p$. Differentiating equality (4.2) with respect to α and evaluating at $\alpha=0$ gives

$$(4.5) \quad X_p A + A^T X_p = -E_2.$$

We have assumed, however, that A has pure imaginary or zero eigenvalues. Let $i\beta \in \sigma(A)$, and denote by $y \in \mathbb{C}^n$ the corresponding eigenvector. Then we have $Ay = i\beta y$, $y^* A^T = -i\beta y^*$, and (4.5) implies $-y^* E_2 y = 0$, a contradiction with the negative definiteness of E_2 . \square

Using Theorem 4.1, we can now describe the Lipschitz stable real symmetric solutions of (3.1).

THEOREM 4.2. *Assume $(A, C, D) \in \mathcal{C}_n$, and assume that the matrix*

$$M = \begin{bmatrix} A & D \\ C & -A^T \end{bmatrix}$$

has no pure imaginary or zero eigenvalues. Then the following statements are equivalent for a real symmetric solution X of

$$(4.6) \quad XDX + XA + A^T X - C = 0:$$

- (i) X is conditionally Lipschitz stable;
- (ii) X is unconditionally Lipschitz stable;
- (iii) $\sigma(A + DX) \cap \sigma(-A^T - XD) = \emptyset$.

Proof. If condition (iii) holds, then the corresponding M -invariant J -Lagrangian subspace $\mathcal{N} = \text{Im} \begin{bmatrix} I \\ X \end{bmatrix}$ is spectral (i.e., sum of root subspaces). Then, clearly, \mathcal{N} is unconditionally Lipschitz stable as an M -invariant J -Lagrangian subspace (even as an M -invariant subspace); so (ii) follows. The implication (ii) \Rightarrow (i) is evident. Finally, assume (i). Let X_0 be the maximal real symmetric solution of (4.6). We have

$$\begin{bmatrix} I & 0 \\ -X_0 & I \end{bmatrix} M \begin{bmatrix} I & 0 \\ X_0 & I \end{bmatrix} = \begin{bmatrix} A + DX_0 & D \\ 0 & -(A + DX_0)^T \end{bmatrix};$$

so without loss of generality we can (and will) assume that $C = 0$ and the spectrum of A lies in the open left half-plane (as $\sigma(M) \cap i\mathbb{R} = \emptyset$, we must have also $\sigma(A) \cap i\mathbb{R} = \emptyset$). Arguing by contradiction, let $\lambda_0 \in \sigma(A + DX) \cap \sigma(-A^T - XD)$, and assume (without loss of generality) that λ_0 lies in the open left half-plane. The corresponding M -invariant subspace $\mathcal{N} = \text{Im} \begin{bmatrix} I \\ X \end{bmatrix}$ is then not a spectral subspace. Let $\mathcal{N}_0 = \text{Im} \begin{bmatrix} I \\ 0 \end{bmatrix}$ and let $P: \mathbb{R}^{2n} \rightarrow \mathbb{R}^n$ be the projection on the first n coordinates (so $P \begin{bmatrix} x \\ y \end{bmatrix} = x$, where $x, y \in \mathbb{R}^n$). The subspace $\mathcal{H} = P(\mathcal{N} \cap \mathcal{N}_0) \subset \mathbb{R}^n$ is clearly A -invariant; moreover, \mathcal{H} is not a spectral subspace for A . Consequently (see [KMR] and [GLR5]), \mathcal{H} is not Lipschitz stable as an A -invariant subspace. So there exists a sequence of real matrices $\{A_p\}_{p=1}^\infty$ such that $A_p \rightarrow A$ as $p \rightarrow \infty$ but

$$(4.6) \quad \lim_{p \rightarrow \infty} \frac{\text{dist}(\mathcal{H}, \text{Inv}(A_p))}{\|A_p - A\|} = \infty,$$

where $\text{Inv}(A_p)$ stands for the set of all real A_p -invariant subspaces, and

$$\text{dist}(\mathcal{H}, \text{Inv}(A_p)) = \inf_{\mathcal{L} \in \text{Inv}(A_p)} \Theta(\mathcal{H}, \mathcal{L})$$

is the distance from \mathcal{H} to $\text{Inv}(A_p)$ in the gap metric. For p large enough, we have $(A_p, 0, D) \in \mathcal{C}_n$, but (4.6) implies

$$\lim_{p \rightarrow \infty} \frac{\text{dist}(X, R(A_p))}{\|A_p - A\|} = \infty,$$

where

$$\text{dist}(X, R(A_p)) = \inf_{Y \in R(A_p)} \|X - Y\|,$$

and $R(A_p)$ stands for the (nonempty) set of real symmetric solutions of the equation $YDY + YA_p + A_p^T Y = 0$. We have obtained a contradiction to the conditional Lipschitz stability of X . \square

The last part of the proof of Theorem 4.2 uses similar ideas to those in the proof of Theorem 4.5 in [RR2].

We conclude this section with an improved version of Theorem 4.9 in [RR2] concerning Lipschitz stable complex Hermitian solutions of the Riccati equation with complex coefficients

$$(4.7) \quad XDX + XA + A^*X - C = 0,$$

where A , D , and C are complex $n \times n$ matrices with $D = D^* \geq 0$, $C = C^*$, and (A, D) controllable. The definitions of conditionally and unconditionally Lipschitz stable Hermitian solutions X of (4.7) are given analogously to the real case (allowing perturbed equations with complex coefficients as well).

THEOREM 4.3. *Assume that $D = D^* \geq 0$, $C = C^*$, and (A, D) controllable.*

There exists an (un)conditionally Lipschitz stable Hermitian solution X of (4.7) if and only if the matrix

$$\begin{bmatrix} A & D \\ C & -A^* \end{bmatrix}$$

has no pure imaginary or zero eigenvalues. In this case a Hermitian solution X of (4.7) is (un)conditionally Lipschitz stable if and only if

$$\sigma(A + DX) \cap \sigma(-A^* - XD) = \emptyset.$$

The proof of Theorem 4.3 is obtained in the same way as the proofs of Theorems 4.1 and 4.2.

5. Further properties of stable solutions of the Riccati equation. In this section we indicate some properties of the stable solutions that follow from Theorems 3.2, 3.3, 4.1, and 4.2. Given the algebraic Riccati equation

$$(5.1) \quad XDX + XA + A^T X - C = 0,$$

with $(A, C, D) \in \mathcal{C}_n$. Denote by \mathcal{S}_{cs} , \mathcal{S}_{us} , and \mathcal{S}_{LS} the classes of conditionally stable, unconditionally stable, and Lipschitz stable real symmetric solutions of (5.1), respectively (by Theorems 4.1 and 4.2, the classes of conditionally and unconditionally Lipschitz stable real symmetric solutions coincide).

COROLLARY 5.1. *We have*

$$\mathcal{S}_{Ls} \subset \mathcal{S}_{us} \subset \mathcal{S}_{cs}.$$

Furthermore, assuming the set of real symmetric solutions of (5.1) is nonempty, we have $\mathcal{S}_{us} = \mathcal{S}_{cs}$ if and only if the matrix

$$M = \begin{bmatrix} A & D \\ C & -A^T \end{bmatrix}$$

has no pure imaginary or zero eigenvalues. The equality $\mathcal{S}_{Ls} = \mathcal{S}_{us}$ holds if and only if every nonreal nonpure imaginary eigenvalue of M with geometric multiplicity 1 has algebraic multiplicity 1, and every nonzero real eigenvalue of M with geometric multiplicity 1 has algebraic multiplicity at most 2.

Next, we show that the stable solutions (with the exception of conditionally stable ones) are persistent in the sense that any sufficiently close solution of a nearby equation is again stable.

COROLLARY 5.2. *Let X_0 belong to one of the classes \mathcal{S}_{Ls} and \mathcal{S}_{us} . Then there exists $\varepsilon > 0$ such that every solution Y_0 of*

$$YD'Y + YA' + A'^TY - C' = 0$$

belongs to the same stability class, provided $(A', C', D') \in \mathcal{C}_n$ and

$$\|A' - A\| + \|C' - C\| + \|D' - D\| + \|Y_0 - X_0\| < \varepsilon.$$

For the conditional stability, this corollary is generally false, as the following example shows.

Example 5.1. Let

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & -1 \end{bmatrix}, \quad C = 0,$$

and D any 4×4 positive semidefinite Hermitian matrix such that (A, D) is controllable and the $(2, 4)$ entry of D is nonzero. Then the zero solution of $XD'X + XA + A^TX = 0$ is conditionally stable by Theorem 3.2. However, for real $\varepsilon \neq 0$ such that $|\varepsilon|$ is small, the matrix

$$M(\varepsilon) = \begin{bmatrix} A(\varepsilon) & D \\ 0 & -A(\varepsilon)^T \end{bmatrix},$$

where

$$A(\varepsilon) = \begin{bmatrix} 1 & 1 \\ 0 & 1 + \varepsilon \end{bmatrix} \oplus \begin{bmatrix} -1 & -1 \\ 0 & -1 - \varepsilon \end{bmatrix}$$

has geometric multiplicity 1 and algebraic multiplicity 2 corresponding to the eigenvalue 1. By the same Theorem 3.2 the zero solution of $XD'X + XA(\varepsilon) + A(\varepsilon)^TX = 0$ is not conditionally stable (for real $\varepsilon \neq 0$ close to zero).

Our last observation relates the notions of stability with isolatedness of the solutions. A real symmetric solution X_0 of (5.1) is called *isolated* if there is no other real symmetric solution in a sufficiently small neighborhood of X_0 , in other words, if there is $\varepsilon > 0$ such that any real symmetric solution $X \neq X_0$ of (5.1) (if such exists) satisfies the inequality $\|X - X_0\| \geq \varepsilon$. It is easy to see (via the isolatedness properties of M -invariant J -Lagrangian subspaces, see [RR1]) that every conditionally stable real symmetric solution is isolated. The converse is generally not true, as the following example shows.

Example 5.2. Let $A(\varepsilon)$, C , D be as in Example 5.1. Then for fixed real $\varepsilon \neq 0$ with $|\varepsilon|$ small, the zero solution of $XD\dot{X} + XA(\varepsilon) + A(\varepsilon)^T X = 0$ is isolated but not conditionally stable.

The following result describes the situation where every isolated solution is conditionally stable.

COROLLARY 5.3. *Every isolated real symmetric solution of (3.1) is conditionally stable if and only if every real nonzero eigenvalue of M of geometric multiplicity 1 has odd algebraic multiplicity.*

The proof of this corollary follows by combining Theorem 3.2 and the description of isolated real invariant subspaces for real matrices (see [BGK]). Analogous statements concerning the case where every isolated solution is unconditionally stable, or Lipschitz stable, can be given as well. We omit these statements.

6. Disconjugacy of Hamiltonian systems. Consider the linear Hamiltonian equation

$$(6.1) \quad Jx' = Hx,$$

where

$$J = \begin{bmatrix} 0 & -I_n \\ I_n & 0 \end{bmatrix}, \quad H = \begin{bmatrix} A & B^T \\ B & C \end{bmatrix}$$

are constant real matrices and $x = x(t)$, $-\infty < t < \infty$, is an unknown $2n$ -dimensional real vector function. Equation (6.1) is called *disconjugate* if it has no nontrivial solutions $x(t) = \begin{bmatrix} y(t) \\ z(t) \end{bmatrix}$ such that the vector $y(t)$ has zeros at two distinct points. We have the following result (see [C]).

THEOREM 6.1. *Assume that $C = C^T$ is positive semidefinite, $A = A^T$, and (C, B) is controllable. Then (6.1) is disconjugate if and only if the Riccati equation*

$$(6.2) \quad XCX + B^T X + XB + A = 0$$

has real symmetric solutions X .

We say that (6.1) is *stably disconjugate* if there exists $\varepsilon > 0$ such that all equations

$$(6.3) \quad Jx' = \tilde{H}x$$

with real matrices

$$\tilde{H} = \begin{bmatrix} \tilde{A} & \tilde{B}^T \\ \tilde{B} & \tilde{C}^T \end{bmatrix}$$

satisfying $\tilde{C} = \tilde{C}^T$ positive semidefinite, $\tilde{A} = \tilde{A}^T$, and $\|\tilde{A} - A\| + \|\tilde{B} - B\| + \|\tilde{C} - C\| < \varepsilon$ are disconjugate. The criterion for stable disconjugacy is given in the following theorem.

THEOREM 6.2. *Equation (6.1) is stably disconjugate if and only if the matrix*

$$(6.4) \quad \begin{bmatrix} B & C \\ -A & -B^T \end{bmatrix}$$

has no pure imaginary or zero eigenvalues.

Proof. Assume that (6.3) has no pure imaginary eigenvalues. By Theorem 2.2, (6.2) has real symmetric solutions, and this is also true for (6.3) if $\varepsilon > 0$ is small enough. By Theorem 6.1, (6.1) is stably disconjugate.

Conversely, assume that (6.4) has pure imaginary or zero eigenvalues, and that (6.1) is disconjugate. We will show that it is not stably disconjugate. To this end (in view of Theorem 6.1), it is sufficient to exhibit for every $\varepsilon > 0$ real matrices \tilde{A} , \tilde{C} , and \tilde{B} such that $\tilde{C} = \tilde{C}^T$ is positive semidefinite, $\tilde{A} = \tilde{A}^T$, $\|\tilde{A} - A\| + \|\tilde{B} - B\| + \|\tilde{C} - C\| < \varepsilon$, and the equation $X\tilde{C}X + \tilde{B}^TX + X\tilde{B} + \tilde{A} = 0$ has no real symmetric solutions. This can be done by using the same perturbations as in the proof of the unconditional stability part of Lemma 3.1. \square

The theorem can also be easily obtained from Corollary 3 in [F2].

REFERENCES

- [BGK] H. BART, I. GOHBERG, AND M. A. KAASHOEK, *Minimal Factorizations of Matrix and Operator Functions*, Birkhäuser-Verlag, Basel, 1979.
- [BC] J. A. BALL AND N. COHEN, *Sensitivity minimization in H^∞ norm: parametrization of all suboptimal solutions*, Internat. J. Control, 46 (1987), pp. 785–816.
- [B] H. BAUMGÄRTEL, *Analytic Perturbation Theory for Matrices and Operators*, Birkhäuser-Verlag, Basel, 1985.
- [Br] R. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.
- [BM] A. BUNSE-GERSTNER AND V. MEHRMANN, *A symplectic QR-like algorithm for the solution of the real algebraic Riccati equations*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 1104–1113.
- [Bu] R. J. BUCY, *Structural stability for the Riccati equation*, SIAM J. Control Optim., 13 (1975), pp. 749–752.
- [By] R. BYERS, *Solving the algebraic Riccati equation with matrix sign function*, Linear Algebra Appl., 85 (1987), pp. 267–279.
- [C] W. A. COPPEL, *Disconjugacy*, Lecture Notes in Math., 220, Springer-Verlag, Berlin, Heidelberg, New York, 1971.
- [Cu] A. N. ČURILOV, *The frequency theorem and the Lur'e equation*, Sibirsk. Math. Zh., 20 (1979), pp. 600–611. (In Russian.)
- [CD] R. CUSHMAN AND J. J. DUISTERMAAT, *The behaviour of the index of a periodic linear Hamiltonian system under iteration*, Adv. Math., 23 (1977), pp. 1–21.
- [D] D. F. DELCHAMPS, *A note on the analyticity of the Riccati metric*, Lectures in Appl. Math., 18, C. J. Byrnes and C. F. Martin, eds., pp. 37–41, AMS, Providence, RI, 1980.
- [DPWZ] D. Z. DJOKOVIC, J. POTERA, P. WINTERNITZ, AND H. ZASSENHAUS, *Normal forms of elements of classical real and complex Lie and Jordan algebras*, J. Math. Phys., 24 (1983), pp. 1363–1374.
- [F] L. E. FAIBUSOVICH, *Algebraic Riccati equation and symplectic algebra*, Internat. J. Control, 43 (1986), pp. 781–792.
- [F2] ———, *Matrix Riccati inequality: existence of solutions*, Systems Control Lett., 9 (1987), pp. 59–65.
- [GGLD] K. GLOVER, M. GREEN, D. LIMEBEER, AND J. DOYLE, *A J -spectral factorization approach to H^∞ control*, preprint.
- [GLR1] I. GOHBERG, P. LANCASTER, AND L. RODMAN, *Spectral analysis of selfadjoint matrix polynomials*, Ann. Math., 112 (1980), pp. 33–71.
- [GLR2] ———, *Matrix Polynomials*, Academic Press, New York, 1982.
- [GLR3] ———, *Perturbation of H -selfadjoint matrices, with applications to differential equations*, Integral Equations Operator Theory, 5 (1982), pp. 718–757.
- [GLR4] ———, *Matrices and Indefinite Scalar Products*, Birkhäuser-Verlag, Basel, 1983.

- [GLR5] I. GOHBERG, P. LANCASTER, AND L. RODMAN, *Invariant Subspaces of Matrices with Applications*, John Wiley, New York, 1986.
- [K] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
- [KS] H. KWAKERNAK AND R. SIVAN, *Linear Optimal Control Systems*, Wiley-Interscience, New York, 1972.
- [KMR] M. A. KAASHOEK, C. V. M. VAN DE MEE, AND L. RODMAN, *Analytic operator functions with compact spectrum, II. Spectral pairs and factorization*, Integral Equations Operator Theory, 5 (1982) pp. 791–827.
- [LR] P. LANCASTER AND L. RODMAN, *Existence and uniqueness theorem for the algebraic Riccati equation*, Internat. J. Control, 32 (1980), pp. 285–309.
- [L] A. J. LAUB, *A Schur method for solving the algebraic Riccati equations*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 913–925.
- [RR1] A. C. M. RAN AND L. RODMAN, *Stability of invariant Lagrangian subspaces I*, Oper. Theory: Adv. Appl., 32, I. Gohberg, ed., Birkhäuser-Verlag, Basel, 1988, pp. 181–228.
- [RR2] ———, *Stability of invariant maximal semidefinite subspaces II, applications: selfadjoint rational matrix functions, algebraic Riccati equation*, Linear Algebra Appl., 63 (1984), pp. 133–173.
- [RR3] ———, *Stability of neutral invariant subspaces in indefinite inner products and stable symmetric factorizations*, Integral Equations Operator Theory, 6 (1983), pp. 536–571.
- [RR4] ———, *On parameter dependence of solutions of algebraic Riccati equations*, Math. Control Signals Systems, 1 (1988), pp. 269–284.
- [RV] A. C. M. RAN AND R. VREUGDENHIL, *Existence and comparison theorems for algebraic Riccati equations for continuous and discrete-time systems*, Linear Algebra Appl., 99 (1988), pp. 63–83.
- [Ro1] L. RODMAN, *Maximal invariant neutral subspaces and application to the algebraic Riccati equation*, Manuscripta Math., 43 (1983), pp. 1–12.
- [Ro2] ———, *On extremal solutions of the algebraic Riccati equation*, Lectures in Applied Math., 18, C. I. Byrne and C. F. Martin, eds., pp. 311–327, AMS, Providence, RI, 1980.
- [S] M. A. SHAYMAN, *Geometry of the algebraic Riccati equations. Parts I and II*, SIAM J. Control Optim., 21 (1983), pp. 374–394, 395–409.
- [ST] J. M. SOETHOUDT AND H. L. TRENTelman, *The regular indefinite linear quadratic problem with linear endpoint constraints*, Systems Control Lett., 12 (1989), pp. 23–31.
- [Th] R. C. THOMPSON, *Pencils of complex and real symmetric and skew matrices*, Linear Algebra Appl., 147 (1991), pp. 323–371.
- [T] H. J. TRENTelman, *The regular free-endpoint linear quadratic problem with indefinite cost*, SIAM J. Control Optim., 27 (1989), pp. 27–42.
- [ZK] K. ZHOU AND P. P. KHARGONEKAR, *An algebraic Riccati equation approach to H^∞ optimization*, Systems Control Lett., 11 (1988), pp. 85–92.

SUFFICIENCY CRITERIA VIA FOCAL POINTS AND VIA COUPLED POINTS*

VERA ZEIDAN†

This paper is dedicated to the memory of W. T. Reid.

Abstract. Considered in this paper is a general quadratic functional $J(\eta)$ with constraints of the form $\eta(a) = D\eta(b) = 0$, where D is an $r \times n$ -matrix ($r \leq n$) and of full rank. It is shown that the nonexistence of focal points to b in $[a, b]$ is equivalent to the existence of a solution to the corresponding Riccati equation with associated boundary conditions. Thus, it is equivalent to the positivity of $J(\eta)$. This result generalizes W. A. Coppel, *Proceedings of the Royal Society of Edinburgh*, 73A, 18, 1974-1975, pp. 271-289; V. B. Haas, *Systems and Control Letters*, 5 (1984), North-Holland, Amsterdam, pp. 55-57; and W. T. Reid, Academic Press, New York, 1972, in which either $D = I$ or 0 is assumed. Moreover, it is proven that the nonexistence of "coupled points with a " in $(a, b]$ is also equivalent to the positivity of $J(\eta)$, proving that for this problem, the notion of a "coupled point with a " is the one searched for to extend that of a "conjugate point to a " from $D = I$ to a general D . Each of these conditions is proved to be sufficient for optimality in the nonlinear calculus of variations problem with fixed initial state but variable final endpoints.

Key words. calculus of variations, focal points, coupled points, Riccati equations, sufficient conditions

AMS(MOS) subject classification. 49B36

1. Introduction. Consider the following functional:

$$J(\eta) = \frac{1}{2} \eta(b)^T \Gamma \eta(b) + \frac{1}{2} \int_a^b \{ \eta^T(s) P(s) \eta(s) + 2 \dot{\eta}^T(s) Q(s) \eta(s) + \dot{\eta}^T(s) R(s) \dot{\eta}(s) \} ds,$$

where $\eta(\cdot)$ is absolutely continuous $\eta: [a, b] \rightarrow \mathbb{R}^n$ and satisfies

$$\eta(a) = 0, \quad D\eta(b) = 0;$$

and $P(\cdot)$, $Q(\cdot)$, and $R(\cdot)$ are given $n \times n$ -matrix functions such that, for all $t \in [a, b]$, $P(t)$ and $R(t)$ are symmetric. The functions P , Q , and R are essentially bounded on $[a, b]$. The matrix Γ is symmetric, and $D \in M_{r \times n}$ ($r \leq n$) is of full rank.

Set

$$\mathcal{D} := D^T (DD^T)^{-1} D;$$

then \mathcal{D} is a *projection*, symmetric, $\mathcal{D}(I - \mathcal{D}) = 0$, and $\mathcal{D}\alpha = 0$ if and only if $D\alpha = 0$. Thus, without loss of generality, the boundary condition $D\eta(b) = 0$ can be replaced by $\mathcal{D}\eta(b) = 0$. Furthermore, due to the boundary conditions on $\eta(b)$, Γ in $J(\eta)$ can be assumed of the form $(I - \mathcal{D})\Gamma(I - \mathcal{D})$.

The Euler-Lagrange equation associated with the problem is

$$(1.1) \quad \frac{d}{dt} [R(t) \dot{\eta}(t) + Q(t) \eta(t)] = Q^T(t) \dot{\eta}(t) + P(t) \eta(t), \quad t \in [a, b] \text{ a.e.}$$

Assume throughout the paper the strengthened Legendre condition, i.e., $R(t) \geq \beta I$, for $t \in [a, b]$ almost everywhere, and for some $\beta > 0$. Then, following Cesari [3], by a solution of (1.1) we mean a function $\eta(\cdot)$ absolutely continuous (AC) for which there

* Received by the editors September 11, 1989, accepted for publication (in revised form) November 27, 1990. The author thanks Consiglio Nazionale delle Ricerche (Italy) and the Natural Sciences and Engineering Research Council of Canada, whose support made this research possible.

† Department of Mathematics, Michigan State University, East Lansing, Michigan 48824.

exists a function $\xi(\cdot) \in AC$ satisfying for $t \in [a, b]$ almost everywhere the *Jacobi system*

$$(1.2) \quad \begin{aligned} \dot{\eta}(t) &= A(t)\eta(t) + B(t)\xi(t), \\ \dot{\xi}(t) &= C(t)\eta(t) - A^T(t)\xi(t), \end{aligned}$$

where

$$A = -R^{-1}Q, \quad B = R^{-1}, \quad C = P - Q^T R^{-1}Q.$$

Since the initial condition of η is fixed, corresponding to \mathbf{b} there is the notion of a *focal point* introduced for the case of smooth data by Kneser in 1898 (see [1] and [16]). Such a point $c \in [a, b]$ is characterized by the existence of a nonzero solution (η, ξ) of (1.2) satisfying

$$(1.3) \quad \eta(c) = \mathcal{D}\eta(b) = 0 \quad \text{and} \quad \langle \eta(b), \xi(b) + \hat{\Gamma}\eta(b) \rangle = 0,$$

where $\hat{\Gamma} = (I - \mathcal{D})\Gamma(I - \mathcal{D})$.

Sufficient conditions for the positivity of $J(\eta)$ are known in terms of the corresponding Riccati equation. See, for instance, [2], where the result is proved for the more general situation, that is, the linear optimal regulator problem. The question of relating focal point theory *directly* to the Riccati equation with a corresponding boundary condition is of theoretical and practical interest. In addition to providing us with a sufficiency criterion in terms of a first-order linear system of ordinary differential equations (1.2), it allows us to see how we can *explicitly* construct a solution to the Riccati equation. Much work has been done in that direction but only for the special case when $\mathcal{D} = 0$ or $\mathcal{D} = I$. In fact, it is shown in [5], [8], and [12] that in those cases the nonexistence of focal points (when $\mathcal{D} = I$ focal points are conjugate points) to \mathbf{b} in $[a, b)$ is equivalent to the existence of a solution to the corresponding Riccati equation and, hence, to the positivity of $J(\eta)$. This is proved by constructing a “conjoined basis” to the Jacobi system in terms of which the solution for the Riccati equation is obtained. On the other hand, it is a well-known fact (when $\mathcal{D} = 0$ see [10] and [12], and for a general \mathcal{D} , see [20]) that the nonexistence of focal points \mathbf{b} in (a, b) is a necessary condition for the nonnegativity of $J(\eta)$. Thus, for $\mathcal{D} = I$ or 0, the gap between necessary and sufficient conditions is as small as possible.

In [7] there is a transformation from a variable endpoint problem to one with a free endpoint. Hence, the Riccati solution to the original problem is given in terms of the one for the free endpoint case. This idea would not work here since $P(b)$ is undefined through that transformation, and thus the boundary condition in (3) of Theorem 2.1 cannot be recaptured.

Recently, in collaboration with Zezza, the author developed in [17]–[19] the notion of a coupled point that is an extension of the focal point to the case of variable endpoint(s). Since the final endpoint of η in the above problem is varying, we have here the notion of a point coupled with \mathbf{a} . To such a point c , corresponds a nonzero solution (η, ξ) of the Jacobi system satisfying

$$\begin{aligned} \eta(\cdot) &\not\equiv \eta(c) \quad \text{on} \quad [c, b] \quad (\text{drop if } c = b) \\ \eta(a) &= \mathcal{D}\eta(c) = 0 \quad \text{and} \quad \left\langle \eta(c), \xi(c) + \hat{\Gamma}\eta(c) + \int_c^b P(s) ds \eta(c) \right\rangle = 0, \end{aligned}$$

where $\hat{\Gamma} = (I - \mathcal{D})\Gamma(I - \mathcal{D})$.

It is shown in [18] and [19] that the nonexistence of points in (a, b) coupled with \mathbf{a} is necessary for the nonnegativity of $J(\eta)$ subject to $\eta(a) = \mathcal{D}\eta(b) = 0$.

This paper focuses on two issues. First, we show that for a general \mathcal{D} , the nonexistence of focal points to \mathbf{b} in $[a, b)$ is equivalent to the existence of a solution

to the corresponding Riccati equation and boundary condition, and hence it is equivalent to the strict positivity of $J(\eta)$ over $\eta(a) = \mathcal{D}\eta(b) = 0$. The proof we employ is of a constructive nature. We show how the nonexistence of focal points to \mathbf{b} in $[a, b]$ produces an *explicit formula* for the solution of the Riccati equation with associated boundary conditions. Thus, we extend the result in [5], [8], and [15] to a general case, i.e., when the right endpoint is restricted to belong to any linear subspace $\mathcal{L} = \ker \mathcal{D}$ of the state space. The second goal is to show that the definition of a coupled point is, in fact, the right one that extends the notion of a focal point to the variable endpoint(s) problem. To accomplish this aim, it is only natural to consider the above problem, that is, the problem with one endpoint *fixed* and the other *varying* on a linear subspace. Then we show that the nonexistence of coupled points to \mathbf{a} also allows us to construct a solution to the corresponding Riccati equation with associated boundary condition, and hence is equivalent to the strict positivity of $J(\eta)$ over $\eta(a) = \mathcal{D}\eta(b) = 0$. This also means that the nonexistence of focal points to \mathbf{b} in $[a, b]$ is equivalent to the nonexistence of coupled points with \mathbf{a} in $(a, b]$. In the last section, we show that the nonexistence of focal points to \mathbf{b} in $[a, b]$, or the nonexistence of coupled points with \mathbf{a} in $(a, b]$ is sufficient for optimality in the nonlinear calculus of variations problem with fixed initial and variable final endpoints. Finally, we provide a numerical example to illustrate the utility of our results.

2. Sufficiency and focal points. Consider the quadratic functional $J(\eta)$ given in the Introduction. The following definition can be found in [1] and [16].

DEFINITION 2.1. A point \mathbf{c} is focal to \mathbf{b} if there exists a nonzero solution (η, ξ) of the Jacobi system

$$(2.1) \quad \begin{aligned} \dot{\eta} &= A\eta + B\xi, \\ \dot{\xi} &= C\eta - A^T\xi, \end{aligned}$$

where $A = -R^{-1}Q$, $B = R^{-1}$, $C = P - Q^TR^{-1}Q$, with

$$(2.2) \quad \eta(c) = \mathcal{D}\eta(b) = 0 \quad \text{and} \quad \langle \eta(b), \xi(b) + \hat{\Gamma}\eta(b) \rangle = 0,$$

where $\hat{\Gamma} = (I - \mathcal{D})\Gamma(I - \mathcal{D})$.

The main goal of this section is to establish the equivalence between the positivity of $J(\eta)$ over $\eta: \eta(a) = \mathcal{D}\eta(b) = 0$ and the nonexistence of focal points to \mathbf{b} in $[a, b]$. This is done by constructing a certain “conjoined basis” (U, V) of (2.1) having a certain property that leads to the construction of a solution to the corresponding Riccati equation.

Following Reid [15], we make the following definition.

DEFINITION 2.2. A pair (U, V) of $n \times n$ -matrix functions on $[a, b]$ is a conjoined basis of the Jacobi system (2.1) if (U, V) solves the Jacobi matrix system, and the columns (U_i, V_i) of (U, V) form a set of linearly independent solutions of (2.1) such that

$$U^TV = V^TU.$$

The following theorem forms the main result of this section.

THEOREM 2.1. *The following are equivalent.*

- (1) *There exists no point in $[a, b]$ focal to \mathbf{b} ,*
- (2) *There exists a conjoined basis (U, V) to (2.1) such that $\det U \neq 0$ on $[a, b]$ and*

$$(I - \mathcal{D})[V(b) + \hat{\Gamma}U(b)] = 0,$$

(3) *There exists a Lipschitz symmetric matrix function $W(\cdot):[a, b] \rightarrow M_{n \times n}$ solution of*

$$\dot{W} + WBW + A^T W + WA - C = 0 \quad \text{for } t \in [a, b] \text{ a.e.,}$$

with

$$(I - \mathcal{D})[W(b) + \hat{\Gamma}] = 0,$$

$$(4) \quad J(\eta) \geq \varepsilon \int_a^b |\dot{\eta}(t)|^2 dt \quad \forall \eta: \eta(a) - D\eta(b) = 0, \quad \text{for some } \varepsilon > 0.$$

Remark. When $\mathcal{D} = I$ or $\mathcal{D} = 0$, Theorem 2.1 reduces to well-known results by Reid [15, Chap. IV, Thm. 7.1]; respectively [15, Chap. IV, Thm. 7.3]. Thus, our result here extends that of Reid to a general setting. Reid has also tackled the case where subspace constraints at both endpoints are allowed. In Chapter IV of [15, Thm. 7.4] he established the relationship between the positivity of the quadratic functional and the corresponding Riccati equation. Here, in Theorem 2.1, we extend his result to also include the focal point theory for the case where the initial endpoint is zero and the final endpoint is constrained to a subspace.

Proof of Theorem 2.1. We will show the following implications: (2) \Rightarrow (3) \Rightarrow (4) \Rightarrow (1) \Rightarrow (2).

(2) \Rightarrow (3): Let (U, V) be the conjoined basis in (2). Since U is invertible on $[a, b]$, define $W = VU^{-1}$. Using $U^T V = V^T U$, we obtain the symmetry of W . It is easily shown that W satisfies the Riccati equation of (3) and $(I - \mathcal{D})[W(b) + \hat{\Gamma}] = 0$.

(3) \Rightarrow (4): Let W be the Lipschitz symmetric function satisfying condition (3). Since $R(t) \geq \beta I$, for $t \in [a, b]$ almost everywhere, then for $0 < \varepsilon \leq \beta/2$, $R_\varepsilon(t) := R(t) - \varepsilon I \geq (\beta/2)I$ for almost all t in $[a, b]$. Consider the Riccati equation in condition (3) with B replaced by $B_\varepsilon = R_\varepsilon^{-1}$. Since $R_\varepsilon \geq (\beta/2)I$ almost everywhere, then $R_\varepsilon^{-1} \in L^\infty[a, b]$ and, hence, by the embedding theorem of differential equations or the continuous dependence on a parameter (see, e.g., Theorem 4.1 in the Appendix of [9]), there exists a Lipschitz symmetric matrix function W_ε satisfying

$$\dot{W}_\varepsilon + W_\varepsilon B_\varepsilon W_\varepsilon + A_\varepsilon^T W_\varepsilon + W_\varepsilon A_\varepsilon - C_\varepsilon = 0 \quad \text{a.e.}$$

with

$$W_\varepsilon(b) = W(b) + \varepsilon I,$$

where $B_\varepsilon = R_\varepsilon^{-1}$, $A_\varepsilon = -R_\varepsilon^{-1}Q$, $C_\varepsilon = P - Q^T R_\varepsilon^{-1}Q$. Let $\eta \neq 0$ be any absolutely continuous function satisfying $\eta(a) = \mathcal{D}\eta(b) = 0$. Then

$$0 = \frac{1}{2} \eta^T(b) W_\varepsilon(b) \eta(b) - \frac{1}{2} \int_a^b \frac{d}{dt} \{ \eta^T(t) W_\varepsilon(t) \eta(t) \} dt$$

and hence

$$\begin{aligned} J(\eta) &= \frac{1}{2} \eta(b)^T (W_\varepsilon(b) + \Gamma) \eta(b) + \frac{1}{2} \int_a^b \{ \eta^T(t) (P(t) - \dot{W}_\varepsilon(t)) \eta(t) \\ &\quad + 2\dot{\eta}^T(t) (Q(t) - W_\varepsilon(t)) \eta(t) + \dot{\eta}^T(t) R(t) \dot{\eta}(t) \} dt \\ &= \frac{1}{2} \eta^T(b) (I - \mathcal{D}) (W_\varepsilon(b) + \Gamma) (I - \mathcal{D}) \eta(b) \\ &\quad + \frac{1}{2} \int_a^b [\eta^T (Q^T - W_\varepsilon) + \dot{\eta}^T R_\varepsilon] R_\varepsilon^{-1} [(Q - W_\varepsilon) \eta + R_\varepsilon \dot{\eta}] dt \\ &\quad + \varepsilon \int_a^b |\dot{\eta}(t)|^2 dt. \end{aligned}$$

Thus, if $J_\varepsilon(\eta)$ is the second term of the last equation, then $J_\varepsilon(\eta) \geq 0$ and

$$J_\varepsilon(\eta) = 0 \quad \text{if and only if} \quad \dot{\eta} = -R_\varepsilon^{-1}(Q - W_\varepsilon)\eta.$$

Since $\eta(a) = 0$, it follows that $J_\varepsilon(\eta)$ is positive definite. Since $(I - \mathcal{D})(W_\varepsilon(b) + \hat{\Gamma}) = \varepsilon(I - \mathcal{D})$, condition (4) holds.

(4) \Rightarrow (1): If there exists a point $c \in [a, b]$ focal to b then there exists $(\eta, \xi) \neq 0$ satisfying (2.1) and (2.2). Define

$$\bar{\eta}(s) = \begin{cases} \eta(s) & \text{on } [c, b] \\ 0 & \text{on } [a, c], \end{cases}$$

then

$$\begin{aligned} J(\bar{\eta}) &= \frac{1}{2} \eta^T(b) \Gamma \eta(b) + \frac{1}{2} \int_c^b \{ \dot{\xi} \cdot \eta + \dot{\eta} \cdot \xi \} dt \\ &= \frac{1}{2} \eta^T(b) \hat{\Gamma} \eta(b) + \frac{1}{2} \xi \cdot \eta|_c^b \\ &= 0, \end{aligned}$$

and hence, using (4), $\bar{\eta} \equiv 0$. Since B is invertible, (2.1) yields $\xi \equiv 0$ on $[c, b]$. Therefore, $(\eta, \xi) \equiv 0$ and we have a contradiction, proving that (1) holds.

Let us define (U_a, V_a) and (U_b, V_b) to be the solutions to the Jacobi matrix system

$$\begin{aligned} \dot{U} &= AU + BV, \\ \dot{V} &= CU - A^T V, \end{aligned} \tag{2.3}$$

with boundary conditions, respectively,

$$\begin{aligned} U_a(a) &= 0, & U_b(b) &= I - \mathcal{D}, \\ V_a(a) &= I, & V_b(b) &= -\hat{\Gamma} - \mathcal{D}. \end{aligned} \quad \text{and}$$

It is easy to see that (U_a, V_a) and (U_b, V_b) satisfy $d/dt[U^T V - V^T U] = 0$ for t in $[a, b]$ almost everywhere. Thus, by using the boundary conditions on (U_a, V_a) and (U_b, V_b) and the fact that $\hat{\Gamma} = (I - \mathcal{D})\Gamma(I - \mathcal{D})$, it results that $U_\alpha^T V_\alpha = V_\alpha^T U_\alpha$, $\alpha = a, b$.

To prove that (1) \Rightarrow (2) the following lemmas will be needed.

LEMMA 2.1. For all $t \in [a, b]$,

$$V_a^T(t) U_b(t) - U_a^T(t) V_b(t) = \text{constant} = U_b(a). \tag{2.4}$$

If, in addition, we assume condition (1) of Theorem 2.1, then $\det U_b(t) \neq 0 \quad \forall t \in [a, b]$ and $\det U_a(b) \neq 0$.

Proof of Lemma 2.1. Using system (2.3) for (U_a, V_a) and (U_b, V_b) , we obtain

$$\frac{d}{dt} [V_a^T(t) U_b(t) - U_a^T(t) V_b(t)] = 0$$

and hence

$$V_a^T(t) U_b(t) - U_a^T(t) V_b(t) = \text{constant} = M.$$

Evaluating $t = a$, we get that $M = U_b(a)$.

If there exists $c \in [a, b]$ and $\alpha (\neq 0) \in \mathbb{R}^n$ such that $U_b(c)\alpha = 0$, define

$$(\eta(t), \xi(t)) = (U_b(t)\alpha, V_b(t)\alpha).$$

Thus, (η, ξ) and c satisfy the Jacobi system (2.1) and (2.2). Moreover, $(\eta, \xi) \neq 0$, since otherwise $(\eta(b), \xi(b)) = ((I - \mathcal{D})\alpha, (-\hat{\Gamma} - \mathcal{D})\alpha) = 0$ and hence, $(I - \mathcal{D})\alpha = 0$ and $\mathcal{D}\alpha = 0$. Therefore, $\alpha = 0$, contradicting $\alpha \neq 0$. Thus, c is focal to b , but this contradicts (1). We have proved that $\det U_b(t) \neq 0$ for all $t \in [a, b]$.

Now, if $\det U_a(b) = 0$, then $U_a(b)\alpha = 0$ for some $\alpha \neq 0$, $\alpha \in \mathbb{R}^n$. Define $(\eta(t), \xi(t)) = (U_a(t)\alpha, V_a(t)\alpha)$. It follows that (2.1) holds, and $\eta(a) = \eta(b) = 0$; i.e., (2.2) is also satisfied for $c = a$. Moreover, $(\eta(a), \xi(a)) = (0, \alpha) \neq 0$ yields $(\eta, \xi) \neq 0$. Therefore, a is focal to b , and a contradiction with (1) is obtained. \square

LEMMA 2.2. *Let*

$$(2.5) \quad \begin{aligned} U(t) &= U_a(t)U_b^{T^{-1}}(a)\mathcal{D} + U_b(t), \\ V(t) &= V_a(t)U_b^{T^{-1}}(a)\mathcal{D} + V_b(t), \end{aligned}$$

then (U, V) satisfies (2.3), $U^T(t)V(t) = V^T(t)U(t)$, $(I - \mathcal{D})(V(b) + \hat{\Gamma}U(b)) = 0$, $\det U(a) \neq 0$ and $\det U(b) \neq 0$.

Proof of Lemma 2.2. Using the definition of (U_a, V_a) , (U_b, V_b) , (2.4), and the invertibility of $U_b(a)$, it results that

$$\begin{aligned} U^T(t)V(t) - V^T(t)U(t) &= \mathcal{D}U_b^{-1}(a)(U_a^T(t)V_b(t) - V_a^T(t)U_b(t)) \\ &\quad + (U_b^T(t)V_a(t) - V_b^T(t)U_a(t))U_b^{T^{-1}}(a)\mathcal{D} \\ &= -\mathcal{D} + \mathcal{D} \\ &= 0. \end{aligned}$$

Since (U_a, V_a) and (U_b, V_b) satisfy (2.3), then so does (U, V) defined by (2.5). Let us calculate

$$\begin{aligned} (I - \mathcal{D})(V(b) + \hat{\Gamma}U(b)) &= (I - \mathcal{D})V_a(b)U_b^{T^{-1}}(a)\mathcal{D} - \hat{\Gamma} \\ &\quad + (I - \mathcal{D})\hat{\Gamma}U_a(b)U_b^{T^{-1}}(a)\mathcal{D} + \hat{\Gamma}. \end{aligned}$$

But, from (2.4) evaluated at $t = b$ we obtain

$$(I - \mathcal{D})V_a(b) + \hat{\Gamma}U_a(b) = (I - \mathcal{D})U_b^T(a);$$

therefore,

$$(I - \mathcal{D})(V(b) + \hat{\Gamma}U(b)) = 0.$$

Note that $U(a) = U_b(a)$, which is invertible. Now we show that $U(b) = U_a(b)U_b^{T^{-1}}(a)\mathcal{D} + (I - \mathcal{D})$ is invertible. If there exists $\alpha (\neq 0) \in \mathbb{R}^n$ with

$$(2.6) \quad U_a(b)U_b^{T^{-1}}(a)\mathcal{D}\alpha = -(I - \mathcal{D})\alpha,$$

then, from (2.4) at $t = b$ we have

$$(I - \mathcal{D})V_a(b) + \hat{\Gamma}U_a(b) + \mathcal{D}U_a(b) = U_b^T(a)$$

and thus

$$\mathcal{D}U_a(b) = \mathcal{D}U_b^T(a),$$

which yields

$$\mathcal{D}U_a(b)U_b^{T^{-1}}(a) = \mathcal{D}.$$

Using this in (2.6), we get $\mathcal{D}\alpha = 0$, but from (2.6) we obtain that $(I - \mathcal{D})\alpha = 0$. Hence $\alpha = 0$ and a contradiction results. Therefore $U(b)$ is invertible. \square

Let us return to the proof of Theorem 2.1.

(1) \Rightarrow (2): From Lemma 2.1 we have $U_b(a)$ invertible. Let (U, V) be defined by (2.5). Using Lemma 2.2, it remains to show that U is invertible on (a, b) . If this is false, then there exists $c \in (a, b)$ and $\alpha \neq 0$ with $U(c)\alpha = 0$, that is,

$$(2.7) \quad U_a(c)U_b^{T^{-1}}(a)\mathcal{D}\alpha = -U_b(c)\alpha.$$

Define

$$(2.8) \quad (\eta(t), \xi(t)) = \begin{cases} (U_a(t)U_b^{T^{-1}}(a)\mathcal{D}\alpha, V_a(t)U_b^{T^{-1}}(a)\mathcal{D}\alpha) & t \in [a, c] \\ -(U_b(t)\alpha, V_b(t)\alpha) & t \in [c, b]. \end{cases}$$

It is clear that η is absolutely continuous with $\eta(a) = \mathcal{D}\eta(b) = 0$. Let us compute $J(\eta)$ by taking into account that (U_a, V_a) and (U_b, V_b) satisfy (2.3). It follows that

$$\begin{aligned} J(\eta) &= \frac{1}{2} \alpha^T \hat{\Gamma} \alpha + \int_a^c \frac{d}{dt} [\alpha^T \mathcal{D} U_b^{-1}(a) V_a^T(t) U_a(t) U_b^{T^{-1}}(a) \mathcal{D} \alpha] dt \\ &\quad + \int_c^b \frac{d}{dt} [\alpha^T U_b^T(t) V_b(t) \alpha] dt. \end{aligned}$$

Using (2.7) we obtain

$$J(\eta) = \frac{1}{2} \alpha^T [-\mathcal{D} U_b^{-1}(a) V_a^T(c) U_b(c) + \mathcal{D} U_b^{-1}(a) U_a^T(c) V_b(c)] \alpha$$

and hence from (2.4) we have

$$(2.9) \quad J(\eta) = -\frac{1}{2} \alpha^T \mathcal{D} \alpha \leq 0.$$

Set

$$h(t) = \begin{cases} U_b^{-1}(t) U_a(t) U_b^{T^{-1}}(a) \mathcal{D} \alpha & t \in [a, c] \\ -\alpha & t \in [c, b]. \end{cases}$$

From (2.7) it results that $h(\cdot)$ is absolutely continuous and $\eta(t) = U_b(t)h(t)$. From the corollary on p. 138 of [15] and the fact that $\eta(a) = 0$, it follows that

$$\begin{aligned} J(\eta) &= \frac{1}{2} \alpha^T \hat{\Gamma} \alpha + \frac{1}{2} h^T(t) U_b^T(t) V^T(t) h(t) \Big|_a^b \\ &\quad + \frac{1}{2} \int_a^b \dot{h}^T(t) U^T(t) R(t) U(t) h(t) dt \\ &= \frac{1}{2} \int_a^b \dot{h}^T(t) U^T(t) R(t) U(t) h(t) dt \geq 0, \end{aligned}$$

since $R(t) > 0$ for all t . Thus, (2.9) gives $\mathcal{D}\alpha = 0$. But from (2.7) and the invertibility of $U_b(c)$ for $c \in [a, b]$, we obtain $\alpha = 0$, which is a contradiction. \square

3. Sufficiency and coupled points. Consider here the same quadratic form as in § 2. In [17]–[19], the notion of coupled points was introduced. It is an extension of that of focal (and conjugate) points to the case of variable endpoint(s). Since the final endpoint $\eta(b)$ is varying in our problem, we have the notion of points coupled with **a**. It has been shown in the above-mentioned references that the nonexistence of points in (a, b) coupled with **a** is a necessary condition for the nonnegativity of the functional

$J(\eta)$ over the constraints $\eta(a) = \mathcal{D}\eta(b) = 0$. It is an important question whether the notion of coupled points exactly replaces in the variable endpoint(s) problem that of focal or conjugate points known for the fixed endpoint(s) case. In this section, we answer this question for the problem studied in the previous sections. We show that the nonexistence of points in $(a, b]$ coupled with \mathbf{a} is equivalent to the positivity of the functional $J(\eta)$ over $\eta(a) = 0 = \mathcal{D}\eta(b)$. This result, combined with that of the previous section, tells us that the nonexistence of points focal to \mathbf{b} in $[a, b]$ is in fact equivalent to the nonexistence of points coupled with \mathbf{a} in $(a, b]$. The approach we use here is analogous to that in § 2. However, the main difficulty resides in showing that the nonexistence of points in $(a, b]$ coupled with b implies condition (2) of Theorem 2.1.

DEFINITION 3.1. A point $c \in (a, b]$ is coupled with \mathbf{a} if there exists a pair $(\eta, \xi) \neq 0$ satisfying the Jacobi system (2.1) with $\eta(\cdot) \neq \eta(c)$ on $[c, b]$ (drop if $c = b$),

$$(3.1) \quad \eta(a) = \mathcal{D}\eta(c) = 0 \quad \text{and} \quad \left\langle \eta(c), \xi(c) + \hat{\Gamma}\eta(c) + \int_c^b P(s) ds \eta(c) \right\rangle = 0,$$

where $\hat{\Gamma} = (I - \mathcal{D})\Gamma(I - \mathcal{D})$.

Remark. When $\mathcal{D} = I$, condition (3.1) reduces to $\eta(a) = \eta(c) = 0$. Moreover, in this case, $\eta(\cdot) \neq \eta(c)$ on $[c, b]$ is trivially satisfied since otherwise $(\eta, \xi) \equiv (0, 0)$ on $[a, b]$. Thus, when $\mathcal{D} \neq I$ and $\eta(c) \neq 0$, the last of (3.1) says that in the definition of a coupled point a certain penalty term $\int_c^b P(s) ds \eta(c)$ should be added to the transversality condition at c .

The following result is the objective of this section.

THEOREM 3.1. *Condition (1) in Theorem 2.1 can be replaced by*

(i) *there exists no point in $(a, b]$ coupled with \mathbf{a} ,*

which is equivalent to

(ii) *there is no $c \in (a, b]$ for which there exists $(\eta, \xi) \neq 0$ satisfying (2.1) on $[a, b]$ and (3.1).*

To prove the theorem we need to study two issues. The first concerns the impact of the strengthened Legendre condition on the problem with “penalized” functional $\tilde{J}([a, t], \eta)$ (see below). Then, we derive important properties of condition (i) of Theorem 3.1.

For $t \in [a, b]$, define

$$\begin{aligned} \tilde{J}([a, t], \eta) := & \frac{1}{2} \eta^T(t) \left(\Gamma + \int_t^b P(s) ds \right) \eta(t) \\ & + \frac{1}{2} \int_a^t \{ \eta^T(s) P(s) \eta(s) + 2 \dot{\eta}^T(s) Q(s) \eta(s) \\ & + \dot{\eta}^T(s) R(s) \dot{\eta}(s) \} ds, \end{aligned}$$

where $\eta: [a, b] \rightarrow \mathbb{R}^n$ is absolutely continuous and satisfies

$$\eta(a) = \mathcal{D}\eta(t) = 0.$$

LEMMA 3.1. *The strengthened Legendre condition yields the existence of $t_0 \in (a, b]$ such that*

$$\tilde{J}([a, t_0], \eta) > 0 \quad \forall \eta \neq 0: \eta(a) = \mathcal{D}\eta(t_0) = 0.$$

Proof of Lemma 3.1. Let β be the positive number satisfying the strengthened Legendre condition. The essential boundedness of $P(\cdot)$ and $Q(\cdot)$ yields the existence of $M > 0$ such that, for $t \in [a, b]$ almost everywhere,

$$\|P(t)\| \leq M, \quad \|Q(t)\| \leq M, \quad \text{for all } t, \quad \left\| \int_a^b P(s) ds \right\| \leq M, \quad \text{and } \|\hat{\Gamma}\| \leq M,$$

where “ $\|\cdot\|$ ” is the matrix 2-norm. Choose $t_0 \in (a, b]$ close enough to a such that

$$-M(t_0 - a)^2 - 4M(t_0 - a) + \beta > 0.$$

Let $\eta(\cdot) \neq 0$: $\eta(a) = \mathcal{D}\eta(t_0) = 0$. We will show that $\tilde{J}([a, t_0], \eta) > 0$. In fact, from the Cauchy-Schwartz inequality it follows that, for all $s \in [a, t_0]$,

$$|\eta(s)| \leq \int_a^{t_0} |\dot{\eta}(\tau)| d\tau \leq (t_0 - a)^{1/2} \left(\int_a^{t_0} |\dot{\eta}(\tau)|^2 d\tau \right)^{1/2}.$$

Thus,

$$\begin{aligned} & \left| \eta^T(t_0) \left(\hat{\Gamma} + \int_{t_0}^b P(s) ds \right) \eta(t_0) + \int_a^{t_0} \{ \eta^T(s) P(s) \eta(s) + 2\dot{\eta}^T(s) Q(s) \eta(s) \} ds \right| \\ & \leq \{ 2M(t_0 - a) + M(t_0 - a)^2 + 2M(t_0 - a) \} \int_a^{t_0} |\dot{\eta}(s)|^2 ds. \end{aligned}$$

Then,

$$\tilde{J}([a, t_0], \eta) > \{-4M(t_0 - a) - M(t_0 - a)^2 + \beta\} \int_a^{t_0} |\dot{\eta}(s)|^2 ds > 0. \quad \square$$

Let (U_a, V_a) be the solution of (2.3) satisfying, as in § 2, $U_a(a) = 0$, and $V_a(a) = I$.

COROLLARY 3.1. *Under the conditions of Lemma 3.1 we have $\det U_a(t_0) \neq 0$, where t_0 is the value in Lemma 3.1, and, for all $\beta \neq 0$: $\mathcal{D}\beta = 0$,*

$$\beta^T \left[V_a(t_0) U_a^{-1}(t_0) + \hat{\Gamma} + \int_{t_0}^b P(s) ds \right] \beta > 0.$$

Proof. If for $\alpha \neq 0$ $U_a(t_0)\alpha = 0$, define on $[a, t_0]$ the functions $(\eta(t), \xi(t)) = (U_a(t)\alpha, V_a(t)\alpha)$. Then (η, ξ) satisfy the Jacobi system (2.1) and $(\eta, \xi) \neq 0$, since $\xi(a) = \alpha \neq 0$. Moreover, $\eta(a) = \eta(t_0) = 0$ and hence, using the Jacobi system,

$$\tilde{J}([a, t_0], \eta) = \eta(t) \cdot \xi(t)|_a^{t_0} = 0.$$

This contradicts Lemma 3.1.

Now, if there exists $\beta \neq 0$: $\mathcal{D}\beta = 0$ and $\beta^T [V_a(t_0) U_a^{-1}(t_0) + \hat{\Gamma} + \int_{t_0}^b P(s) ds] \beta \leq 0$, define on $[a, t_0]$, $(\eta(t), \xi(t)) = (U_a(t) U_a^{-1}(t_0) \beta, V_a(t) U_a^{-1}(t_0) \beta)$. Since $\xi(a) = U_a^{-1}(t_0) \beta \neq 0$, then $(\eta, \xi) \neq 0$. Also, $\eta(a) = \mathcal{D}\eta(t_0) = 0$ and

$$\begin{aligned} \tilde{J}([a, t_0], \eta) &= \frac{1}{2} \beta^T \left[\Gamma + \int_{t_0}^b P(s) ds \right] \beta + \eta(t) \cdot \xi(t)|_a^{t_0} \\ &= \frac{1}{2} \beta^T \left[\hat{\Gamma} + \int_{t_0}^b P(s) ds + V_a(t_0) U_a^{-1}(t_0) \right] \beta \\ &\leq 0. \end{aligned}$$

This contradicts Lemma 3.1. \square

Let us now study the properties of the nonexistence of coupled points.

LEMMA 3.2. *Assume that \mathbf{b} is not coupled with \mathbf{a} , then in the definition of c coupled with a (Definition 3.1) the condition $\eta(\cdot) \neq \eta(c)$ on $[c, b]$ can be eliminated.*

Proof of Lemma 3.2. The result will be proved by contradiction. If there exist $c \in (a, b]$ and $(\eta, \xi) \neq 0$ satisfying (2.1) and (3.1) with $\eta(\cdot) \equiv \eta(c)$ on $[c, b]$, then on $[c, b]$, (2.1) yields

$$\xi(t) = Q(t)\eta(c)$$

and

$$\dot{\xi}(t) = P(t)\eta(c).$$

Thus

$$\xi(b) = \xi(c) + \int_c^b P(s) ds \eta(c)$$

and hence (3.1) becomes

$$\eta(a) = 0 = \mathcal{D}\eta(b) \quad \text{and} \quad \langle \eta(b), \xi(b) + \hat{\Gamma}\eta(b) \rangle = 0,$$

that is, b is coupled with a . Therefore, we obtain a contradiction. \square

Let (U_b, V_b) be the pair as in § 2, that is, the solution of (2.3) with $U_b(b) = I - \mathcal{D}$ and $V_b(b) = -\hat{\Gamma} - \mathcal{D}$.

LEMMA 3.3. *If there exists no $c \in (a, b]$ coupled with a then $\det U_a(t) \neq 0$ on $(a, b]$ and $\det U_b(a) \neq 0$.*

Proof of Lemma 3.3. If for some $c \in (a, b]$ and $\alpha \neq 0$ we have $U_a(c)\alpha = 0$, then $(\eta(t), \xi(t)) = (U_a(t)\alpha, V_a(t)\alpha) \neq 0$ satisfies (2.1) and $\eta(a) = \eta(c) = 0$. This contradicts the hypothesis that no $c \in (a, b]$ is coupled, and hence conjugate to a .

If $U_b(a)\alpha = 0$ for $\alpha \neq 0$, then

$$(\eta(t), \xi(t)) = (U_b(t)\alpha, V_b(t)\alpha)$$

satisfies $\eta(a) = \mathcal{D}\eta(b) = 0$ and $\langle \eta(b), \xi(b) + \hat{\Gamma}\eta(b) \rangle = 0$. Moreover, $(\eta(b), \xi(b)) = ((I - \mathcal{D})\alpha, -\hat{\Gamma}\alpha - \mathcal{D}\alpha) \neq 0$ (otherwise $\alpha = 0$). Thus, we have b coupled with a , and hence a contradiction follows. \square

The following results say that if (i) of Theorem 3.1 holds, then in the definition of a coupled point not only $\eta(\cdot) \neq \eta(c)$ on $[c, b]$ can be removed but the last equality in (3.1) can be replaced by “ \leq .”

LEMMA 3.4. *Assume that there are no points in $(a, b]$ coupled with a , then there exist no $c \in (a, b]$ and $(\eta, \xi) \neq 0$ solution of (2.1) with*

$$\eta(a) = \mathcal{D}\eta(c) = 0 \quad \text{and} \quad \left\langle \eta(c), \xi(c) + \hat{\Gamma}\eta(c) + \int_c^b P(s) ds \eta(c) \right\rangle \leq 0.$$

Proof of Lemma 3.4. From Definition 3.1 and the assumption of Lemma 3.4, we only need to show that the strict inequality above cannot happen. Let us argue by contradiction. Assume there exist $c \in (a, b]$ and $(\eta, \xi) \neq 0$ satisfying system (2.1), $\eta(a) = \mathcal{D}\eta(c) = 0$ and $\langle \eta(c), \xi(c) + \hat{\Gamma}\eta(c) + \int_c^b P(s) ds \eta(c) \rangle < 0$. We can easily see that since $\eta(a) = 0$ and (η, ξ) is nonzero and satisfies (2.1), there exists $\alpha_0 \neq 0$ such that $(\eta(t), \xi(t)) = (U_a(t)\alpha_0, V_a(t)\alpha_0)$. Thus, also $\mathcal{D}U_a(c)\alpha_0 = 0$ and

$$\alpha_0^T U_a^T(c) \left[V_a(c) + \hat{\Gamma}U_a(c) + \int_c^b P(s) ds U_a(c) \right] \alpha_0 < 0.$$

Set $\beta_0 = U_a(c)\alpha_0$. Then from Lemma 3.3 we have $\beta_0 \neq 0$, also $\mathcal{D}\beta_0 = 0$ and

$$\beta_0^T \left[V_a(c)U_a^{-1}(c) + \hat{\Gamma} + \int_c^b P(s) ds \right] \beta_0 < 0.$$

Define

$$\mathcal{A}(t) := V_a(t)U_a^{-1}(t) + \hat{\Gamma} + \int_t^b P(s) ds,$$

then $\mathcal{A}(\cdot)$ is continuous on $(a, b]$. From Corollary 3.1, $\beta^T \mathcal{A}(t_0)\beta > 0$ for all $\beta \neq 0$: $\mathcal{D}\beta = 0$. Since, we know $\beta_0 \mathcal{A}(c)\beta_0 < 0$, there exists $\bar{c} \in (a, b]$ such that $\beta_0^T \mathcal{A}(\bar{c})\beta_0 = 0$. Define on $[a, b]$

$$(\eta(t), \xi(t)) = (U_a(t)U_a^{-1}(\bar{c})\beta_0, \quad V_a(t)U_a^{-1}(\bar{c})\beta_0).$$

We have $\eta(a) = \mathcal{D}\eta(\bar{c}) = 0$, $(\eta, \xi) \neq 0$ because $\beta_0 \neq 0$, and $\langle \eta(\bar{c}), \xi(\bar{c}) + \hat{\Gamma}\eta(\bar{c}) + \int_{\bar{c}}^b P(s) ds \eta(\bar{c}) \rangle = \beta_0^T \mathcal{A}(\bar{c})\beta_0 = 0$. Hence, \bar{c} is coupled with a , which is a contradiction. \square

Now we will provide the proof of the main result of this section.

Proof of Theorem 3.1. From Theorem 2.1 and Lemma 3.2, we only need to prove that condition (4) of Theorem 2.1 implies (ii), and (ii) implies condition (2) of Theorem 2.1. Assume condition (4) of Theorem 2.1. If (ii) is false, there exist $(\eta, \xi) \neq 0$ solution of (2.1) and $c \in (a, b]$ such that

$$\eta(a) = \mathcal{D}\eta(c) = 0 \quad \text{and} \quad \left\langle \eta(c), \xi(c) + \hat{\Gamma}\eta(c) + \int_c^b P(s) ds \eta(c) \right\rangle = 0.$$

Define

$$\bar{\eta}(t) := \begin{cases} \eta(t) & \text{on } [a, c], \\ \eta(c) & \text{on } [c, b]; \end{cases}$$

then $\bar{\eta}(c)$ is absolutely continuous, $\bar{\eta}(a) = \mathcal{D}\bar{\eta}(b) = 0$, and

$$\begin{aligned} J(\bar{\eta}) &= \frac{1}{2} \eta^T(c) \left(\hat{\Gamma} + \int_c^b P(s) ds \right) \eta(c) + \frac{1}{2} \eta(t) \cdot \xi(t) \Big|_a^c \\ &= 0. \end{aligned}$$

Also $\bar{\eta} \neq 0$, since otherwise (2.1) gives that $(\eta, \xi) \equiv 0$ on $[a, c]$ and hence on $[a, b]$. Thus, we have a contradiction with condition (4).

Finally, assume that (ii) holds and let us show that condition (2) of Theorem 2.1 is satisfied. From Lemma 3.3 it follows that $U_b(a)$ is invertible. Thus, define (U, V) by (2.5). Using Lemma 2.2, it only remains to show that U is invertible on (a, b) . Suppose not, then there exist $c \in (a, b)$ and $\alpha \neq 0$ such that $U(c)\alpha = 0$, that is, $U_a(c)U_b^{T^{-1}}(a)\mathcal{D}\alpha = -U_b(c)\alpha$, which is (2.7). Define (η, ξ) exactly as in (2.8),

$$(\eta(t), \xi(t)) = \begin{cases} (U_a(t)U_b^{T^{-1}}(a)\mathcal{D}\alpha, V_a(t)U_b^{T^{-1}}(a)\mathcal{D}\alpha) & t \in [a, c], \\ -(U_b(t)\alpha, V_b(t)\alpha) & t \in [c, b]. \end{cases}$$

Then, computations identical to those that led to (2.9) give that $J(\eta) = -\frac{1}{2}\alpha^T \mathcal{D}\alpha \leq 0$. Now, we know by Lemma 3.3 that U_a is invertible on $(a, b]$. Thus, define

$$\gamma(t) = \begin{cases} U_b^{T^{-1}}(a)\mathcal{D}\alpha & t \in [a, c], \\ -U_a^{-1}(t)U_b(t)\alpha & t \in [c, b], \end{cases}$$

then

$$\eta(t) = U_a(t)\gamma(t).$$

Using the corollary on p. 138 of [11] to compute $J(\eta)$, we get that

$$\begin{aligned} J(\eta) &= \frac{1}{2} \eta^T(b) \hat{\Gamma} \eta(b) + \frac{1}{2} \gamma^T(t) V_a^T(t) U_a(t) \gamma(t) \Big|_a^b \\ &\quad + \frac{1}{2} \int_a^b \dot{\gamma}^T(t) U_a^T(t) R(t) U_a(t) \dot{\gamma}(t) dt \\ &= \frac{1}{2} \gamma^T(b) U_a^T(b) [\hat{\Gamma} U_a(b) + V_a(b)] \gamma(b) \\ &\quad + \frac{1}{2} \int_a^b \dot{\gamma}^T(t) U_a^T(t) R(t) U_a(t) \dot{\gamma}(t) dt. \end{aligned}$$

Since (U_a, V_a) is a principal solution to (2.1) with $\eta(a) = 0$, then Lemma 3.4 means that

$$w^T U_a^T(t) \left[V_a(t) + \hat{\Gamma} U_a(t) + \int_t^b P(s) ds U_a(t) \right] w > 0$$

for all $w \neq 0$ and $t \in (a, b]$ such that $\mathcal{D}U_a(t)w = 0$. In particular, $\gamma^T(b) U_a^T(b) \times [V_a(b) + \hat{\Gamma} U_a(b)] \gamma(b) \geq 0$, since $\mathcal{D}U_a(b) \gamma(b) = 0$. Thus, $J(\eta) \geq 0$. We know that $J(\eta) = -\frac{1}{2} \alpha^T \mathcal{D} \alpha \leq 0$, then $J(\eta) = 0$. Thus, $\mathcal{D} \alpha = 0$ and $\gamma(b) = 0$, that is, $\alpha = 0$. Hence a contradiction follows. \square

4. Application to nonlinear problems. In this section we consider the following general problem of calculus of variations with fixed initial state but variable final endpoints:

$$(V) \quad \text{minimize } I(x) := \gamma(x(b)) + \int_a^b L(t, x(t), \dot{x}(t)) dt,$$

over all absolutely continuous functions $x: [a, b] \rightarrow \mathbb{R}^n$ satisfying

$$x(a) = A, \quad \phi(x(b)) = 0,$$

where $A \in \mathbb{R}^n$, $\phi: \mathbb{R}^n \rightarrow \mathbb{R}^r$ ($r \leq n$), $\gamma: \mathbb{R}^n \rightarrow \mathbb{R}$, and $L: [a, b] \times \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$.

In the special case where $\phi(\cdot)$ is affine ($\phi(y) = y - B$) this problem reduces to the *classical* calculus of variations about which intensive literature can be found (see, for instance, [1], [9], [14], and [16]). Recently, the interest in the problem was renewed, due to its close connection to the optimal control problem [16] and to the multiple integral calculus of variations problem [6]. Two of the main questions that are now the center of attention are the existence of Lipschitz solutions to the classical problem (see [4] and the references therein) and the study of the variable endpoint(s) case (see [12]–[14] and [21], and the references therein). As is the case with the conjugate point theory for the classical setting, we would like to obtain for the variable endpoint(s) problem necessary conditions for optimality that are almost sufficient. In [17]–[19], it is shown that in the variable endpoints case the nonexistence of coupled points with \mathbf{b} in (a, b) is necessary for weak local minimality. On the other hand, given that in (V) the left-hand point $x(a)$ is fixed, it is known (see, for instance, [20]) that the nonexistence of focal points to a in (a, b) is necessary. Using the results of previous sections, we show here that each of these necessary conditions can be strengthened to become sufficient for the weak local optimality in (V).

Let $z(\cdot): [a, b] \rightarrow \mathbb{R}^k$ be in $L^\infty[a, b]$. The *tube* of radius ε about $z(\cdot)$ is defined by $T(z(\cdot); \varepsilon) = \{(t, y) \in [a, b] \times \mathbb{R}^k : |y - z(t)| < \varepsilon\}$. A function $y(\cdot)$ is said to be in $T(z(\cdot); \varepsilon)$ if $(t, y(t)) \in T(z(\cdot); \varepsilon)$ for $t \in [a, b]$ almost everywhere.

DEFINITION 4.1. An absolutely continuous function $\hat{x}(\cdot):[a, b] \rightarrow \mathbb{R}^n$ that satisfies the constraints $\hat{x}(a) = A$, $\phi(\hat{x}(b)) = 0$ is said to be *admissible*. An admissible function $\hat{x}(\cdot)$ is a *weak local minimum* for (V) if there exists $\varepsilon > 0$ such that $I(x) \geq I(\hat{x})$ for all admissible $x(\cdot): (x(\cdot), \dot{x}(\cdot)) \in T(\hat{x}, \hat{x}; \varepsilon)$.

Let $\hat{x}(\cdot)$ be admissible and Lipschitz. We make the following nonrestrictive assumptions on the data.

(i) There exists $\varepsilon > 0$ such that $\gamma(\cdot)$ and $\phi(\cdot)$ are C^2 on the ε -neighborhood of $\hat{x}(b)$, and, for almost all $t \in [a, b]$, $L(t, \cdot, \cdot)$ is C^2 on the ε -neighborhood of $(\hat{x}(t), \dot{\hat{x}}(t))$.

(ii) $L(t, x, v)$ and its derivatives in (x, v) up to second order are integrable along $(\hat{x}, \dot{\hat{x}})$.

(iii) $\nabla_{(x,v)}^2 L(t, \cdot, \cdot)$ is continuous at $(\hat{x}, \dot{\hat{x}})$ uniformly in t , and $\hat{L}_{vx}(t) := L_{vx}(t, \hat{x}(t), \dot{\hat{x}}(t))$ is essentially bounded on $[a, b]$.

Given that we are interested in finding a sufficiency criterion for weak local minimality of a Lipschitz candidate \hat{x} , it is natural to assume that \hat{x} satisfies the necessary conditions for optimality: in particular, the Euler-Lagrange equation:

$$(\mathcal{E}) \quad \frac{d}{dt} \hat{L}_v(t) = \hat{L}_x(t) \quad t \in [a, b] \text{ a.e.},$$

the transversality condition:

there exists a vector $l \in \mathbb{R}^r$ such that

$$(\mathcal{T}) \quad -\hat{L}_v(b) = l^T [\nabla \phi(\hat{x}(b))] + \nabla \gamma(\hat{x}(b)),$$

and the Legendre condition:

$$(\mathcal{L}) \quad \hat{L}_{vv}(t) \geq 0 \quad \text{for } t \in [a, b] \text{ a.e.}$$

Define

$$P(s) := \hat{L}_{xx}(s), \quad Q(s) := \hat{L}_{vx}(s), \quad R(s) := \hat{L}_{vv}(s),$$

$$\Gamma = [\nabla^2 \phi(\hat{x}(b))]^T l + \nabla^2 \gamma(\hat{x}(b)),$$

$$D = \nabla \phi(\hat{x}(b)), \quad \mathcal{D} = D^T (DD^T)^{-1} D,$$

and

$$\hat{\Gamma} = (I - \mathcal{D})\Gamma(I - \mathcal{D}).$$

The second variation corresponding to the problem (V) (see, for instance, [20]) is

$$J(\hat{x}; \eta) = \frac{1}{2} \eta(b)^T \Gamma \eta(b) + \frac{1}{2} \int_a^b \{ \eta^T(s) P(s) \eta(s) + 2 \dot{\eta}^T(s) Q(s) \eta(s) + \dot{\eta}^T(s) R(s) \dot{\eta}(s) \} ds,$$

where $\eta(\cdot):[a, b] \rightarrow \mathbb{R}^n$ is absolutely continuous and satisfies $\eta(a) = \mathcal{D}\eta(b) = 0$.

This functional is exactly of the form considered in the previous sections. Thus, as in §§ 2 and 3, we have the notions of focal points to **b** and coupled points with **a**. In terms of these notions, necessary conditions for optimality in (V) were developed in [18] and [19]. The corresponding sufficiency criterion are given by the following theorem, where $(\mathcal{L})'$ denotes the strengthened Legendre condition.

THEOREM 4.1. Let \hat{x} be a Lipschitz admissible function for (V). Assume that \hat{x} satisfies (\mathcal{E}) , (\mathcal{T}) , and $(\mathcal{L})'$. Then \hat{x} provides a strict weak local minimum for (V) if one of the following conditions holds.

(1) There are no points in $[a, b]$ focal to **b**.

- (2) *There are no points in $(a, b]$ coupled with a .*
 (3) *There exists a Lipschitz symmetric matrix function $W(\cdot)$ solution of*

$$\begin{aligned} \dot{W} + WBW + A^T W + WA - C &= 0 \quad t \in [a, b] \text{ a.e.,} \\ (I - \mathcal{D})[W(b) + \hat{\Gamma}] &= 0, \end{aligned}$$

(where A , B , and C are defined as in § 2).

Proof. From Theorems 2.1 and 3.1, each of the above conditions is equivalent to: there exists $\delta > 0$ such that

$$J(\hat{x}; \eta) \geq \delta \int_a^b |\dot{\eta}(t)|^2 dt \quad \forall \eta: \eta(a) = D\eta(b) = 0.$$

Using [10, Thm. 12.2.7], we can find $\alpha > 0$, $\beta > 0$ such that, for all $\eta: \eta(a) = 0$,

$$(4.1) \quad J(\hat{x}; \eta) + \alpha |D\eta(b)|^2 \geq \beta \int_a^b |\dot{\eta}(t)|^2 dt.$$

Define $M = \min \{\beta/8, \beta/(8(b-a)), \beta/(8(b-a)^2)\}$. Since $\phi(\cdot)$ and $\gamma(\cdot)$ are C^2 in the ε -neighborhood of $\hat{x}(b)$, then there exist $K > 0$, $0 < \varepsilon_1 (\leq \varepsilon)$, such that, for all $x: |x - \hat{x}(b)| < \varepsilon_1$,

$$(4.2) \quad \begin{aligned} \|\nabla^2 \phi(x)^T l - \nabla^2 \phi(\hat{x}(b))^T l\| &< M, \\ \|\nabla^2 \gamma(x) - \nabla^2 \gamma(\hat{x}(b))\| &< M, \\ \|\nabla^2 \phi(x)\| &\leq K, \end{aligned}$$

where $\|\cdot\|$ is any matrix norm. Using the fact that $\nabla^2 L(t, \cdot, \cdot)$ is continuous at $(\hat{x}, \dot{\hat{x}})$ uniformly in t , we can find $\varepsilon_2 (\leq \varepsilon_1)$ such that for all absolutely continuous functions $x(\cdot): (x(\cdot), \dot{x}(\cdot)) \in T(\hat{x}, \dot{\hat{x}}; \varepsilon_2)$ we have

$$(4.3) \quad \|\nabla_{(x,v)}^2 L(t, x(t), \dot{x}(t)) - \nabla_{(x,v)}^2 \hat{L}(t)\| < M \quad \text{a.e.}$$

Set $\varepsilon_0 = \min \{\varepsilon_2, \sqrt{\beta/(\alpha K^2(b-a)^3)}\}$, and take $x(\cdot) (\neq \hat{x})$ be any admissible function for (V) with $(x(\cdot), \dot{x}(\cdot)) \in T(\hat{x}, \dot{\hat{x}}; \varepsilon_0)$. We will show that $I(x) > I(\hat{x})$.

Using $\phi(x(b)) = \phi(\hat{x}(b)) = 0$ and Taylor's expansion we get the following:

$$\begin{aligned} I(x) - I(\hat{x}) &= \gamma(x(b)) + \phi(x(b))^T l - (\gamma(\hat{x}(b)) + \phi(\hat{x}(b))^T l) \\ &\quad + \int_a^b \{L(t, x(t), \dot{x}(t)) - \hat{L}(t)\} dt \\ &= [l^T \nabla \phi(\hat{x}(b)) + \nabla \gamma(\hat{x}(b))](x(b) - \hat{x}(b)) \\ &\quad + \int_a^b (\hat{L}_x(t), \hat{L}_v(t)) \begin{pmatrix} x(t) - \hat{x}(t) \\ \dot{x}(t) - \dot{\hat{x}}(t) \end{pmatrix} dt \\ &\quad + \frac{1}{2} (x(b) - \hat{x}(b))^T (\nabla^2 \phi(x')^T l + \nabla^2 \gamma(x'')) (x(b) - \hat{x}(b)) \\ &\quad + \frac{1}{2} \int_a^b (x^T(t) - \hat{x}^T(t), \dot{x}^T(t) - \dot{\hat{x}}^T(t)) \\ &\quad \cdot \nabla^2 L(t, \tilde{x}(t), \dot{\tilde{x}}(t)) \begin{pmatrix} x(t) - \hat{x}(t) \\ \dot{x}(t) - \dot{\hat{x}}(t) \end{pmatrix} dt, \end{aligned}$$

where x' and x'' are between $x(b)$ and $\hat{x}(b)$, and $(\tilde{x}(\cdot), \dot{\tilde{x}}(\cdot))$ is between $(x(\cdot), \dot{x}(\cdot))$ and $(\hat{x}(\cdot), \dot{\hat{x}}(\cdot))$. Integrating by parts the second term and using conditions (\mathcal{E}) and (\mathcal{T}) , we are left with the last two terms. Since $x(a) = \hat{x}(a) = 0$, Hölder's inequality yields that, for all $s \in [a, b]$

$$(4.4) \quad |x(s) - \hat{x}(s)|^2 \leq (b-a) \int_a^b |\dot{x}(t) - \dot{\hat{x}}(t)|^2 dt.$$

Using the first two inequalities of (4.2)–(4.4), and the definition of M , we obtain

$$I(x) - I(\hat{x}) \geq J(\hat{x}; (x - \hat{x})) - \frac{\beta}{4} \int_a^b |\dot{x} - \dot{\hat{x}}|^2 dt,$$

and from (4.1) it follows that

$$I(x) - I(\hat{x}) \geq \frac{3\beta}{4} \int_a^b |\dot{x} - \dot{\hat{x}}|^2 dt - \alpha |D(x(b) - \hat{x}(b))|^2.$$

We also have

$$\begin{aligned} 0 &= \phi(x(b)) - \phi(\hat{x}(b)) \\ &= \nabla \phi(\hat{x}(b))(x(b) - \hat{x}(b)) + \frac{1}{2} (x(b) - \hat{x}(b))^T \nabla^2 \phi(\bar{x})(x(b) - \hat{x}(b)), \end{aligned}$$

and thus,

$$\begin{aligned} |D(x(b) - \hat{x}(b))|^2 &\leq \frac{K^2}{2} |x(b) - \hat{x}(b)|^4 \\ &\leq \frac{K^2}{2} (b-a)^2 \left(\int_a^b |\dot{x} - \dot{\hat{x}}|^2 \right)^2. \end{aligned}$$

Therefore, using $|\dot{x}(t) - \dot{\hat{x}}(t)|^2 < \varepsilon_0^2 \leq \beta / (\alpha K^2 (b-a)^3)$, it results that

$$I(x) - I(\hat{x}) \geq \frac{\beta}{4} \int_a^b |\dot{x} - \dot{\hat{x}}|^2 dt. \quad \square$$

To illustrate the utility of the previous result we consider the following example.

Example. The question is to find a weak local minimum for the problem

$$\begin{aligned} (\bar{V}) \quad \text{Minimize } I(x) &= \frac{1}{2} x_1^2 \left(\frac{\pi}{4} \right) + \frac{1}{2} x_2^2 \left(\frac{\pi}{4} \right) + x_1^3 \left(\frac{\pi}{4} \right) + x_2^5 \left(\frac{\pi}{4} \right) \\ &\quad + \frac{1}{2} \int_0^{\pi/4} \{x_1^2 + x_2^2 - x_1^3 - x_2^3 - x_1 x_2^2 - x_1^2 x_2\} dt \end{aligned}$$

over all absolutely continuous functions $x(\cdot) = \begin{pmatrix} x_1(\cdot) \\ x_2(\cdot) \end{pmatrix}$ satisfying $x_1(0) = x_2(0) = 0$, $x_1(\pi/4) = x_2(\pi/4)$.

Take $\hat{x}(\cdot) = \begin{pmatrix} x_1(\cdot) \\ x_2(\cdot) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$. We can then easily check that $\hat{x}(\cdot)$ satisfies the Euler-Lagrange equation, and the transversality condition for $l = 0$, as well as the strengthened Legendre condition. To prove that $\hat{x}(\cdot)$ is a weak local minimum for (\bar{V}) we will use Theorem 4.1 with condition (2).

For this problem,

$$\begin{aligned} P(s) &\equiv \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \quad Q(s) \equiv 0, \quad R(s) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \\ \Gamma &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \text{and} \quad D = (-1, 1). \end{aligned}$$

The Jacobi system is

$$\begin{aligned}\dot{\eta}_1 &= \xi_1, & \dot{\xi}_1 &= -\eta_1, \\ \dot{\eta}_2 &= \xi_2, & \dot{\xi}_2 &= -\eta_2,\end{aligned}$$

whose solution is

$$(4.5) \quad \begin{aligned}\eta_1(t) &= A \sin t + B \cos t, & \xi_1(t) &= A \cos t - B \sin t, \\ \eta_2(t) &= M \sin t + N \cos t, & \xi_2(t) &= M \cos t - N \sin t.\end{aligned}$$

Let us search for the coupled point with 0 in $(0, \pi/4]$.

From Definition (3.1), $c \in (0, \pi/4]$ is coupled with 0 if there exists $(\eta, \xi) \neq 0$ solving (4.5) with $\eta_1(0) = \eta_2(0) = 0$, $\eta_1(c) = \eta_2(c)$, and $\eta_1(c)[\xi_1(c) + \xi_2(c) + 2\eta_1(c) - 2(\pi/4 - c)\eta_1(c)] = 0$. This is equivalent to $B = N = 0$, $A = M$, and $A \sin c = 0$ or $A[\cos c + \sin c(1 + c - \pi/4)] = 0$. But $A \neq 0$, since otherwise $(\eta, \xi) = (0, 0)$. Thus, we must have $\cos c = -\sin c(1 + c - \pi/4)$. However, $c \in (0, \pi/4]$ and hence the last equation cannot happen. Therefore, there exists no $c \in (0, \pi/4]$ coupled with 0, and by Theorem 4.1, $\hat{x} \equiv 0$ is a weak local minimum for (\bar{V}) .

Acknowledgments. The author thanks the conscientious referees, whose comments helped improve the readability of the paper.

REFERENCES

- [1] G. A. BLISS, *Lectures on the Calculus of Variations*, University of Chicago Press, Chicago, IL, 1946.
- [2] A. E. BRYSON AND Y. C. HO, *Applied Optimal Control*, Blaisdell, Boston, 1969.
- [3] L. CESARI, *Optimization—Theory and Applications*, Springer-Verlag, New York, 1983.
- [4] F. H. CLARKE, *Methods of Dynamic and Nonsmooth Optimization*, Society of Industrial and Applied Mathematics, Philadelphia, PA, 1989.
- [5] W. A. COPPEL, *Linear-quadratic optimal control*, Proc. Roy. Soc. Edinburgh, 73A, 18 (1974/5), pp. 271–289.
- [6] B. DACOROGNA, *Direct Methods in the Calculus of Variations*, Appl. Math. Sci., 78, Springer-Verlag, Berlin, Heidelberg, 1989.
- [7] S. E. DREYFUS, *Control problems with linear dynamics, quadratic criterion, and linear terminal constraints*, IEEE Trans. Automat. Control, AC-12 (1967), pp. 323–324.
- [8] V. B. HAAS, *Linear-quadratic optimal control revisited*, Systems Control Lett., 5 (1984), North-Holland, Amsterdam, pp. 55–57.
- [9] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.
- [10] D. H. JACOBSON, D. H. MARTIN, M. PACHTER, AND T. GEVECI, *Extensions of linear-quadratic control theory*, Lecture Notes in Control and Inform. Sci., No. 27, A. V. Balakrishnan and M. Thoma, eds., Springer-Verlag, Berlin, Heidelberg, New York, 1980.
- [11] V. JURDJEVIC AND J. KOGAN, *Optimality of extremals for linear systems with quadratic costs*, J. Math. Anal. Appl., 143 (1989), pp. 86–108.
- [12] J. KOGAN, *Structure of minimizers in linear-quadratic bolza problems of optimal control*, J. Opt. Theory Appl., Vol. 63, No. 2, 1989.
- [13] J. MAWHIN AND M. WILLEM, *Critical Point Theory and Hamiltonian Systems*, Springer-Verlag, New York, 1989.
- [14] M. MORSE, *Variational Analysis*, John Wiley, New York, 1973.
- [15] W. T. REID, *Riccati Differential Equations*, Academic Press, New York, 1972.
- [16] H. SAGAN, *Introduction to the Calculus of Variations*, McGraw-Hill, New York, 1969.
- [17] V. ZEIDAN AND P. ZEZZA, *An Extension of the Conjugate Point Condition to the Case of Variable End Points*, in Proc. 27th IEEE Conf. Decision Control, 1988, pp. 1187–1191.
- [18] ———, *Variable end-points problems in the calculus of variations: coupled points*, Lecture Notes in Control and Inform. Sci., A. Benoussan and J. L. Lions (INRIA III), eds., Springer-Verlag, Berlin, Heidelberg, 1988, pp. 372–380.

- [19] V. ZEIDAN AND P. ZEZZA, *Coupled points in the calculus of variations and applications to periodic solutions*, Trans. Amer. Math. Soc., 315 (1989), pp. 323–335.
- [20] ———, *Necessary conditions for optimal control problems: conjugate points*, SIAM J. Control Optim., 26 (1988), pp. 592–608.
- [21] E. ZEIDLER, *Nonlinear Functional Analysis and its Applications III: Variational Methods and Optimization*, Springer-Verlag, New York, 1985.

PROPERTIES OF ENERGY-MINIMIZING SEGMENTATIONS*

JAYANT SHAH†

Abstract. In Computer Vision, one approach to segmenting an image consists in minimizing an energy functional that is defined over a set of all possible segmentations in terms of penalties for deviations from ideal properties. Studied here are the smoothing properties of such a formulation defined with two parameters, which are the weights associated with the penalty measures. This paper deals with only the one-dimensional case. It is shown that the effect of these parameters is to set *two* local thresholds, one for the intensity gradient and one for the difference between the maximum and the minimum values of image intensity in a region. If one of the thresholds is not exceeded in a region, the region is regarded as uniform and will not be broken up. Thus, low intensity noise and low gradients are filtered out. Conversely, if the image intensity changes rapidly in a region so that both the thresholds are exceeded, the region will be broken up.

Key words. nonconvex minimization, free boundary problem, image segmentation, piecewise smooth approximations

AMS(MOS) subject classification. 35R35

1. Introduction. In Computer Vision, the segmentation problem is the problem of subdividing an image into regions so that in each region, the image properties are relatively uniform. We have been studying the problem by a variational approach. This approach is motivated in part by occasional failures of traditional methods, which are based on either local edge operators or on histogram partitioning and in part by a desire to integrate the preprocessing and postprocessing steps associated with the traditional methods into one global formulation. The general idea is that we should define an energy functional over a set of all possible segmentations in terms of penalty measures that correspond to various desired properties of a good segmentation. Our approach is a modification of one due to Geman and Geman [5] and subsequently developed by Marroquin [6] and by Blake and Zisserman [2]. The particular functional that we have studied is the following:

$$E(f, B) = \mu^2 \int \int_R (f - g)^2 dx dy + \int \int_{R-B} \|\nabla f\|^2 dx dy - \nu |B|$$

where

R is the image domain,

g is the grey level function, $g: R \rightarrow \mathbf{R}_+$,

B denotes the union of region boundaries; thus B is the segmenting curve,

f is the smoothed image which need not be continuous across B ,

$|B|$ = the length of B ,

μ, ν are the weights.

The problem is to find f and B that minimize $E(f, B)$. While the first term imposes penalty for deviation of f from g , the second term forces f to be as smooth as possible. By minimizing the length of the segmenting curve, the third term tries to avoid segmenting the image into too many regions with wildly zigzagging boundaries. Thus the formulation is designed to find a minimal segmentation such that in every region, the image intensity g is approximately constant. The formulation is minimal in the sense that by dropping any one of the three terms, we get $\inf E = 0$: without the first

* Received by the editors March 12, 1990; accepted for publication (in revised form) January 23, 1991.

† Mathematics Department, Northeastern University, Boston, Massachusetts 02115. This work was supported by National Science Foundation grant IRI-8704467.

term, take $f=0$ and B empty; without the second, take $f=g$ and B empty; without the third, take B to be a fine grid of N horizontal lines and vertical lines, segmenting R into tiny squares and f =average of g in each square.

For a fixed segmenting curve B , let f_B denote the minimizer of $E(f, B)$ with respect to f . In the interior of $R - B$, f_B satisfies the equation $\nabla^2 f_B = \mu^2(f_B - g)$. If B is sufficiently regular, say piecewise C^1 , then f_B satisfies the boundary condition $\partial f_B / \partial n = 0$ along B and along the boundary of R . Thus in each component S of $R - B$, f_B is a smoothed version of g . The amount of smoothing depends on μ . $f_B \rightarrow g$ as $\mu|S| \rightarrow \infty$ and $f_B \rightarrow \bar{g}_S$ as $\mu|S| \rightarrow 0$, where \bar{g}_S is the constant function with value equal to the average of g in S . $1/\mu$ may be thought of as the nominal smoothing radius. For reasonably regular curves B , properties of f_B are well understood, and f_B may be calculated easily by methods such as the finite element method combined with a multigrid relaxation procedure. The real problem is to show the existence and regularity of curves B that minimize $E(f_B, B)$. This is a difficult problem, both theoretically and practically, because of nonlinearity and the existence of many local minima. The questions that arise naturally are

1. Is the problem of minimizing $E(f_B, B)$ with respect to B well posed? In particular, does it have a solution that is not too wild, say, a solution that is piecewise C^2 ?

2. Is there a practical algorithm for minimizing $E(f_B, B)$?

3. Is the formulation well suited for solving vision problems? For example, how is the segmenting curve placed? How does it behave in the presence of noise? How should one choose and vary μ and ν in the context of the vision problem?

All of these questions are still open. We have reported in [7] our initial numerical experiments based on steepest gradient descent. We have also extensively analyzed the limiting cases of the formulation, namely, the limit as $\mu \rightarrow \infty$ and the limit as $\mu \rightarrow 0$ in [8]. We have also discussed the nature of singularities of the segmenting curve B . Asymptotic behavior of the model has also been studied by Richardson in [9]. A very deep analysis of the question of the weak existence of the minimizing curve B has been carried out by Ambrosio [1]; De Giorgi, Carriero, and Leaci [4]; and Dal Maso, Morel, and Solimini [3].

In this paper, we begin to study segmentations obtained when μ and ν are arbitrary. Here, we deal only with the one-dimensional case. That is, we assume that R is one-dimensional and B consists of a set of breakpoints.

The existence of a minimizing segmentation in the one-dimensional case is very easy to see. We note that $\inf_{f,B} E(f, B) \leq E(f_\phi, \phi)$ where f_ϕ is the solution of the Neumann problem with B empty. Therefore, there exists a minimizing sequence $\{f_i, B_i\}$ for $E(f, B)$ such that $|B_i| \leq E(f_\phi, \phi)/\nu$. Hence there exists a subsequence $\{B_{i_k}\}$ converging to B^* , with $|B^*| \leq \inf |B_{i_k}|$. Let f_{B^*} satisfy the equation $f_{B^*}'' = \mu^2(f_{B^*} - g)$ in $R - B^*$ and the boundary condition $f_{B^*}' = 0$ at each breakpoint in B^* and at the endpoints of R . Because the elliptic boundary value problem for f_{B^*} is well posed with respect to deformation of B^* , it follows that $E(f_{B^*}, B^*) \leq \lim E(f_i, B_i)$; hence (f_{B^*}, B^*) minimizes $E(f, B)$.

The interesting question in the one-dimensional case is how μ and ν control the segmentation of R . We show that the role of μ and ν may be interpreted as follows. These parameters set, in effect, *two* local thresholds. The smaller the region, the higher are the thresholds for the region. A region in which

- (i) either the maximum intensity gradient is below its threshold,
- (ii) or the difference between the maximum intensity and the minimum intensity is below its threshold,

is regarded as uniform and will not be broken up. In other words, low intensity noise and low gradients will be filtered out. In a region of R where the intensity changes rapidly such that both the thresholds are exceeded, the region will tend to be broken up such that in each piece, $g_{\max} - g_{\min}$ roughly equals its threshold value. Figure 1 illustrates this behavior.

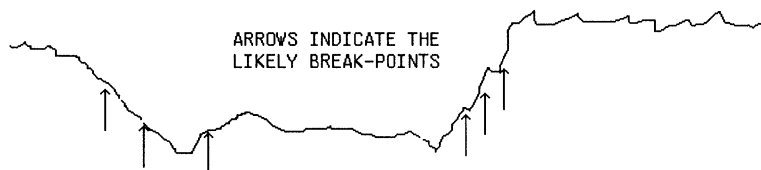


FIG. 1

In § 2, we derive an a priori lower bound on the length of each segment in a minimizing segmentation. From this, we deduce the smoothing properties. We illustrate how the number of segments in a minimizing segmentation varies as we vary μ and ν . In § 3, we derive an upper bound on the segment length to show that the lower bound is qualitatively correct. In the last two sections, we analyze the behavior as $\mu \rightarrow 0$ or $\mu \rightarrow \infty$.

The main technique in deriving these results is obtaining estimates for the reduction in energy due to a single additional cut. Consequently, the upper and lower bounds that we derive hold under much weaker conditions. This allows us to consider segmentations by sets of breakpoints, B , which satisfy conditions

$$(*) \quad \begin{aligned} E(f_B, B) &< E(f_{\tilde{B}}, \tilde{B}) \\ \text{for all } \tilde{B} \subset B \text{ such that } |\tilde{B}| &= |B| - 1 \end{aligned}$$

and

$$(**) \quad \begin{aligned} E(f_B, B) &\leq E(f_{\tilde{B}}, \tilde{B}) \\ \text{for all } \tilde{B} \supset B \text{ such that } |\tilde{B}| &= |B| + 1. \end{aligned}$$

In particular, if B minimizes $E(f_B, B)$, then B satisfies (*) and (**). However, all the conclusions stated above still hold for any B that satisfies (*) and (**), thus indicating that such a set B may already provide an acceptable segmentation.

Notation. For any interval $D \subset R$, define

$$\begin{aligned} \delta_D g &= \max_D g - \min_D g, \\ \text{Lip}_D g &= \begin{cases} \text{Lipschitz constant of } g \text{ in } D & \text{if } g \text{ is Lipschitz,} \\ \infty & \text{otherwise,} \end{cases} \\ \alpha_D &= \max \left\{ \frac{\nu/\mu}{2(\delta_D g)^2}, \frac{\mu\nu}{2(\text{Lip}_D g)^2} \right\}. \end{aligned}$$

2. A lower bound.

THEOREM 1. *Let B be a set of breakpoints satisfying the condition (*). If B is not empty, then the following must hold:*

- (i) $\alpha_R < 1$,
- (ii) for all segments S ,

$$|S| > \frac{1}{\mu} \log \frac{1}{1 - \alpha_R}.$$

Proof. Suppose that B is not empty and satisfied (*). Let S be a segment. Let S' be an adjoining segment. We adopt the notation and the coordinate x as shown in Fig. 2.

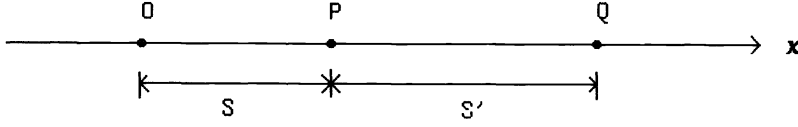


FIG. 2

Let B' be the set of breakpoints obtained by removing the breakpoint at P . Integration by parts (see [8, § 6, Lemma]) gives us

$$0 < E(f_{B'}, B') - E(f_B, B) = f_{B'}'(P)[f_B^+(P) - f_B^-(P)] - \nu$$

where the superscripts $+$ and $-$ refer to the values of f_B on the two sides of a breakpoint. Therefore,

$$\nu < \|f_{B'}'\|_{\infty, S \cup S'} |f_B^+(P) - f_B^-(P)|.$$

We need the following lemma.

LEMMA 1. Let $D = [0, a]$ and $g: D \rightarrow \mathbf{R}$ be a bounded function. Let u satisfy $u'' = \mu^2(u - g)$ and $u'(0) = u'(a) = 0$. Then,

$$g_{\min} \leq u(x) \leq g_{\max},$$

$$|u(x) - g(x)| \leq \min \left\{ \delta_D g, \frac{1}{\mu} \text{Lip}_D g \right\},$$

$$|u'(x)| \leq \min \{ 2\mu \delta_D g, \text{Lip}_D g \} (1 - e^{-\mu x}).$$

Proof of Lemma 1. Extend g to all of \mathbf{R} by successive reflections of D . Let g denote also this extension.

Then,

$$u(x) = \frac{\mu}{2} \int_{-\infty}^{\infty} e^{-\mu r} g(y) dy \quad \text{where } r = |y - x|.$$

Therefore,

$$g_{\min} = \frac{\mu}{2} \int_{-\infty}^{\infty} e^{-\mu r} g_{\min} dy \leq u(x) \leq \frac{\mu}{2} \int_{-\infty}^{\infty} e^{-\mu r} g_{\max} dy = g_{\max}.$$

Moreover,

$$|u(x) - g(x)| \leq \frac{\mu}{2} \int_{-\infty}^{\infty} e^{-\mu r} |g(y) - g(x)| dy \leq \min \left\{ \delta_D g, \frac{1}{\mu} \text{Lip}_D g \right\}.$$

Let $g_c = g - c$, where c is some fixed constant. Since $u'(x)$ remains unchanged, if we add a constant to g ,

$$\begin{aligned} u'(x) &= \frac{\mu}{2} \int_{-\infty}^{\infty} \frac{d}{dx} e^{-\mu r} g_c(y) dy \\ &= -\frac{\mu^2}{2} \int_{-\infty}^x e^{-\mu r} g_c(y) dy + \frac{\mu^2}{2} \int_x^{\infty} e^{-\mu r} g_c(y) dy \\ &= -\frac{\mu^2}{2} \int_{-\infty}^0 e^{-\mu(x-y)} g_c(y) dy - \frac{\mu^2}{2} \int_0^x e^{-\mu(x-y)} g_c(y) dy \\ &\quad + \frac{\mu^2}{2} \int_x^{\infty} e^{-\mu(y-x)} g_c(y) dy. \end{aligned}$$

Since $g(-y) = g(y)$,

$$\begin{aligned} u'(x) &= -\frac{\mu^2}{2} \int_0^\infty e^{-\mu(x+y)} g_c(y) dy - \frac{\mu^2}{2} \int_0^x e^{-\mu(x-y)} g_c(y) dy \\ &\quad + \frac{\mu^2}{2} \int_x^\infty e^{-\mu(y-x)} g_c(y) dy \\ &= -\mu^2 e^{-\mu x} \int_0^x (\cosh \mu y) g_c(y) dy + \mu^2 (\sinh \mu x) \int_x^\infty e^{-\mu y} g_c(y) dy. \end{aligned}$$

Setting $c = (g_{\max} + g_{\min})/2$ so that $|g_c| \leq \frac{1}{2}(\delta_D g)$, we get

$$|u'(x)| \leq \mu(1 - e^{-2\mu x})(\delta_D g) \leq 2\mu(1 - e^{-\mu x})(\delta_D g).$$

Setting $c = g(x)$ so that $|g_c(y)| \leq (\text{Lip}_D g)|y - x|$, we get

$$\begin{aligned} |u'(x)| &\leq \mu^2 (\text{Lip}_D g) \left[e^{-\mu x} \int_0^x (\cosh \mu y)(x - y) dy + (\sinh \mu x) \int_x^\infty e^{-\mu y}(y - x) dy \right] \\ &\leq (1 - e^{-\mu x})(\text{Lip}_D g). \end{aligned} \quad \square$$

We continue now with the proof of the theorem. Applying the lemma to each segment of R , we get

$$|f_B^+(P) - f_B^-(P)| \leq \delta_R g.$$

Also,

$$\begin{aligned} |f_B^+(P) - f_B^-(P)| &= |(f_B^+(P) - g(P)) - (f_B^-(P) - g(P))| \\ &\leq \frac{2}{\mu} \text{Lip}_R g. \end{aligned}$$

Applying the lemma to $S \cup S'$, we get

$$\|f_{B'}\|_{\infty, S \cup S'} \leq \min \{2\mu \delta_R g, \text{Lip}_R g\} (1 - e^{-\mu|S|}).$$

Therefore

$$\begin{aligned} \nu &< 2 \min \left\{ \mu(\delta_R g)^2, \frac{1}{\mu} (\text{Lip}_R g)^2 \right\} (1 - e^{-\mu|S|}) \\ &< 2 \min \left\{ \mu(\delta_R g)^2, \frac{1}{\mu} (\text{Lip}_R g)^2 \right\}. \end{aligned}$$

The theorem follows. \square

Since the smallest segment must have length less than or equal to $|R|/|B|$, part (ii) of Theorem 1 may be restated as follows.

COROLLARY 1. *Let B be a nonempty set of breakpoints satisfying the condition (*). Then,*

$$\nu < 2 \min \left\{ \mu(\delta_R g)^2, \frac{1}{\mu} (\text{Lip}_R g)^2 \right\} (1 - e^{-\mu|R|/|B|}).$$

Thus, the (μ, ν) space is laminated by a series of curves, marking regions corresponding to the number of possible breakpoints in a minimizing segmentation as shown in Fig. 3. An upper bound derived in the next section shows that these curves are qualitatively correct in depicting curves with $|B|$ constant.

To see where the actual breakpoints are likely to occur, we have Corollary 2.

COROLLARY 2. *Let D be a connected interval in R . If*

$$\alpha_D > 1 - e^{-\mu|D|/2},$$

then any set of breakpoints satisfying () cannot have more than two breakpoints in D ; that is, D can contain at most one whole segment.*

Proof. Suppose that there are more than two breakpoints in D . Then, apply the theorem to the union of two adjacent segments contained in D in order to get a contradiction. \square

We may think of Corollary 2 as setting local thresholds (that depend on $|D|$) for $\delta_D g$ and $\text{Lip}_D g$. In particular, if

$$\text{either } \delta_D g < \sqrt{\nu/2\mu} \quad \text{or} \quad \text{Lip}_D g < \sqrt{\mu\nu/2},$$

then there cannot be more than two breakpoints in D . That is, low intensity noise and low gradients are filtered out.

3. An upper bound. We now derive an upper bound on the segment length as a function of μ and ν to show that the lower bound in Theorem 1 is reasonable. An upper bound cannot exist in terms of the global quantities $\delta_R g$ and $\text{Lip}_R g$ used in Theorem 1. To see this, just take g to be a step function. If we keep μ fixed and require a breakpoint at the discontinuity of g , then $\nu \rightarrow 0$ as we move the discontinuity closer and closer to one of the endpoints of R . Therefore, to get an upper bound, we have to make some assumptions regarding the profile of g . It is easy to see that with μ and g fixed and g sufficiently general, we get more and more breakpoints as we decrease ν . Corollary 2 indicates that most of these breakpoints will occur in regions of high gradient. Therefore, we derive an upper bound for segments within which g has high gradient everywhere.

THEOREM 2. *Let B be a set of breakpoints satisfying the condition (**). Suppose that g is C^1 in a segment S and $|g'(x)| \geq c$ for all $x \in S$. Then*

$$\mu|S| \leq \psi^{-1}\left(\frac{\mu\nu}{2c^2}\right),$$

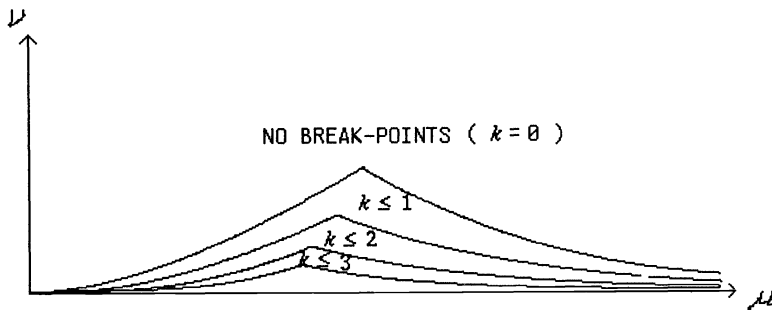


FIG. 3. $k = |B|$.

where ψ is a monotonically increasing function: $\lim_{z \rightarrow \infty} \psi(z) = 1$ and $\psi(z) \approx z^3/32$ for small z .

To prove the theorem, we need Lemma 2.

LEMMA 2. Let $D = [0, a]$ and $g : D \rightarrow \mathbf{R}$ be a C^1 function such that $g'(x) \geq 0$ for all $x \in D$. Let u satisfy $u'' = \mu^2(u - g)$ and $u'(0) = u'(a) = 0$. Then $u'(x) \geq 0$ for all $x \in D$.

Proof. Let $v = u'$. Then, v satisfies $v'' - \mu^2(v - g') = 0$, $v(0) = v(a) = 0$. Therefore, v minimizes

$$U(w) = \mu^2 \int_D (w - g')^2 + \int_D |w'|^2$$

subject to the condition $w(0) = w(a) = 0$. Define

$$\tilde{v}(x) = \begin{cases} v(x) & \text{if } v(x) \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Then, $U(\tilde{v}) \leq U(v)$ and hence $\tilde{v} = v$. Now let

$$u(x) = u(0) + \int_0^x v \, dx \quad \text{where } u(0) = \frac{v'(0)}{\mu^2} + g(0).$$

Proof of Theorem 2. Choose the coordinate x so that the origin is at one of the end points of S and $g'(x) \geq c$. Let g_1 be the linear function with slope c such that $g_1(0) = g(0)$. Write $g = g_1 + \tilde{g}$. In S , let f_1 satisfy $f_1'' = \mu^2(f_1 - g_1)$, $f_1'|_{\partial S} = 0$, and let \tilde{f} satisfy $\tilde{f}'' = \mu^2(\tilde{f} - \tilde{g})$, $\tilde{f}'|_{\partial S} = 0$. Let $\ell = |S|$. Since $g'_1(x) = c$ and $\tilde{g}'(x) \geq 0$, $f'_1(\ell/2) \geq \tilde{f}'(\ell/2) \geq 0$ by Lemma 2. It is easy to construct f_1 explicitly:

$$f_1(x) = g_1(x) - \frac{c \sinh(x - (\ell/2))}{\mu \cosh(\mu\ell/2)}$$

and

$$f'_1(x) = c - \frac{c \cosh(x - (\ell/2))}{\cosh(\mu\ell/2)}.$$

Therefore,

$$f'_1(\ell/2) \geq c \left[1 - \frac{1}{\cosh(\mu\ell/2)} \right].$$

Let B be a set of breakpoints satisfying (**). Consider introduction of an additional breakpoint, P , at the center of the segment S . Let $B' = B \cup \{P\}$. Let $h = f_{B'} - f_B$. Then h satisfies the equation $h'' = \mu^2 h$ in $S - P$ and the boundary condition $h' = 0$ at the endpoints of S , and $h' = -f'_B$ at P . By solving explicitly for h , we easily obtain

$$h^+(\ell/2) = -h^-(\ell/2) = \frac{1}{\mu} f'_B(\ell/2) \coth(\mu\ell/2)$$

Therefore, the change in energy due to the extra breakpoint

$$\begin{aligned} &= E(f_B, B) - E(f_B, B') \\ &= f'_B(P)[f_B^+(P) - f_B^-(P)] - \nu \\ &= f'_B(P)[h^+(P) - h^-(P)] - \nu \\ &= \frac{2}{\mu} [f'_B(\ell/2)]^2 \coth(\mu\ell/2) - \nu \\ &\leq 0 \quad \text{by the condition (**).} \end{aligned}$$

Therefore

$$\begin{aligned} \nu &\geq \frac{2}{\mu} [f'_B(l/2)]^2 \coth(\mu l/2) \\ &\geq \frac{2}{\mu} [f'_1(l/2)]^2 \coth(\mu l/2) \\ &\geq \frac{2}{\mu} c^2 \psi(\mu l), \end{aligned}$$

where

$$\psi(\mu l) = \frac{8 \sinh^4(\mu l/4)}{\sinh(\mu l)}.$$

To compare the lower bound λ , given by Theorem 1, with the upper bound Λ , given by Theorem 2, assume that g is *linear* in a fixed R . Let

$$\theta = \frac{\mu \nu}{2(g')^2}.$$

Suppose that the minimizing B is not empty so that $\theta < 1$. Then,

$$\lambda = \frac{1}{\mu} \varphi^{-1} \left(\theta \max \left\{ \frac{1}{\mu^2 |R|^2}, 1 \right\} \right) \quad \text{where } \varphi(z) = 1 - e^{-z}.$$

We must consider the following two cases.

Case 1. $\mu \Lambda \rightarrow \infty$.

Note that $\mu \Lambda \rightarrow \infty \Leftrightarrow \mu \lambda \rightarrow \infty \Leftrightarrow \theta \rightarrow 1$. Moreover, $\mu \rightarrow \infty$ as $\theta \rightarrow 1$. We have

$$\lim_{\theta \rightarrow 1} \frac{\lambda}{\Lambda} = \lim_{\theta \rightarrow 1} \frac{\varphi^{-1}(\theta)}{\psi^{-1}(\theta)} = \frac{1}{2}.$$

Case 2. $\mu \Lambda \rightarrow 0$ or $\mu \lambda \rightarrow 0$.

$\mu \Lambda \rightarrow 0 \Leftrightarrow \theta \rightarrow 0$ and $\mu \lambda \approx \sqrt[3]{32\theta}$ when θ is small. In order to estimate $\mu \lambda$, we have to consider two subcases.

Case 2a. $\mu |R| \geq 1$. Then $\mu \lambda = \varphi^{-1}(\theta) \approx \theta$ if θ is small. Hence $\Lambda \rightarrow 0 \Leftrightarrow \lambda \rightarrow 0 \Leftrightarrow \theta \rightarrow 0$; however,

$$0 \leq \lim_{\theta \rightarrow 0} \frac{\lambda}{\Lambda} \leq \lim_{\theta \rightarrow 0} \frac{\theta^{2/3}}{\sqrt[3]{32}} = 0.$$

Case 2b. $\mu |R| \leq 1$. Then

$$\mu \lambda = \varphi^{-1} \left(\frac{\theta}{\mu^2 |R|^2} \right) \approx \frac{\theta}{\mu^2 |R|^2} \quad \text{if } \frac{\theta}{\mu^2 |R|^2} \text{ is small.}$$

Hence $\Lambda \rightarrow 0 \Leftrightarrow \lambda \rightarrow 0 \Leftrightarrow \theta/\mu^3 \rightarrow 0$; but,

$$0 \leq \lim_{\theta/\mu^3 \rightarrow 0} \frac{\lambda}{\Lambda} \leq \lim_{\theta/\mu^3 \rightarrow 0} \frac{1}{\sqrt[3]{32}} \left(\frac{\theta}{\mu^3 |R|^3} \right)^{2/3} = 0.$$

Thus, although both bounds tend to zero simultaneously, the lower bound given by Theorem 1 is too low when g is linear and the segment sizes tend to zero. The reason for this is that in the proof of Theorem 1, we had to use the global estimate δ_{Rg} in place of $\delta_{S \cup S'} g$, because the set of breakpoints B need not be maximal among the sets

satisfying the condition (*), and hence $|S'|$ need not be small when $|S|$ is. However, as $\lambda \rightarrow 0$, $\delta_{S \cup S'} g \rightarrow 0$ if B satisfies (**), and $|g'(x)| \geq c > 0$ for all $x \in R$. We can use the upper bound Λ in this special case to estimate $|S \cup S'|$ and thus estimate $\delta_{S \cup S'} g$. Using this estimate, we can obtain a better lower bound, $\lambda_{\#}$ for segmentations that satisfy (*) and (**) such that $\lambda_{\#}/\Lambda$ is uniformly bounded from above and below by positive constants.

4. Behaviour when μ is small. By Theorem 1, in a nontrivial minimizing segmentation,

$$\max \left\{ \frac{\nu/\mu^2}{2(\delta_R g)^2}, \frac{\nu}{2(\text{Lip}_R g)^2} \right\} < \frac{1}{\mu} (1 - e^{-\mu|S|}) < |S| \leq \frac{|R|}{2}.$$

Hence, as $\mu \rightarrow 0$, we must bound ν/μ^2 in order to have nontrivial segmentations. We thus impose the condition

$$\frac{\nu}{\mu^2} = \nu_0 < |R|(\delta_R g)^2$$

when μ is small. Let

$$E_0(B) = \int_R (g - \bar{g}_B)^2 + \nu_0 |B|,$$

where \bar{g}_B is the piecewise constant function that, in each segment, equals the average value of g in that segment. By expressing f and g as cosine series, it is easy to show that

$$(\#) \quad \frac{E_0(B)}{1 + (\mu^2 |R|^2 / \pi^2)} \leq \frac{1}{\mu^2} E(f_B, B) \leq E_0(B).$$

(See [8, § 4] for a similar estimate when R is two-dimensional.) Consequently, we expect the behavior of $E(f_B, B)$ to be controlled by $E_0(B)$ as $\mu \rightarrow 0$. Notice that E_0 has a well-defined minimum. Like Theorem 1, we have the following lemma.

LEMMA 3. *Let B_0 be a nonempty set of breakpoints such that $E_0(B_0) < E_0(B'_0)$ for all $B'_0 \subset B$ with $|B'_0| = |B_0| - 1$. Then, for every segment S ,*

$$|S| > \frac{\nu_0}{(\delta_R g)^2}.$$

Proof. We proceed as in Theorem 1. Consider a segment S . Let S' be an adjoining segment, meeting S at P . Let $s = |S|$ and $s' = |S'|$. Let a , a' and b be the average values of g in S , S' , and $S \cup S'$, respectively. Note that $b = (as + a's')/(s + s')$. Let $B'_0 = B_0 - \{P\}$. Then

$$\begin{aligned} E_0(B'_0) - E_0(B_0) &= \int_{S \cup S'} (g - b)^2 - \int_S (g - a)^2 - \int_{S'} (g - a')^2 - \nu_0 \\ &= \frac{(a - a')^2 ss'}{s + s'} - \nu_0 \\ &\geq 0. \end{aligned}$$

Therefore,

$$\frac{1}{s} < \frac{1}{s} + \frac{1}{s'} = \frac{s + s'}{ss'} \leq \frac{(a - a')^2}{\nu_0} \leq \frac{(\delta_R g)^2}{\nu_0}.$$

As in Corollary 2, if

$$(\delta_D g)^2 < \frac{\nu_0}{|D|}$$

in a connected interval D , then D cannot contain more than one whole segment of the segmentation that minimizes E_0 . Thus, minimizing of E_0 is akin to the schemes that segment images by partitioning histograms.

For each positive integer k , let $D_k \subset \mathbf{R}^k$ be the subspace corresponding to the sets of k breakpoints that result in segments of length greater than $\nu_0/\{2(\partial_R g)^2\}$. $E(f_B, B)$ and $E_0(B)$ achieve their global minima over the space $D = \Pi_{k \leq k_0} D_k$, where k_0 is a fixed integer. To compare the global minima of $E(f_B, B)$ and $E_0(B)$, we may restrict them to D . $E(f_B, B)$ and $E_0(B)$ are continuous over D , where D has the induced topology from $\Pi_{k \leq k_0} \mathbf{R}^k$ with the standard topology. By $(\#)$, $(1/\mu^2)E(f_B, B)$ converges uniformly to $E_0(B)$ over D . From this, it is easy to see Theorem 3.

THEOREM 3. *Let B_0 be a set of breakpoints belonging to D .*

A. Suppose that B_0 minimizes $E_0(B)$ locally. Then, there exist a sequence of sets $\{B_i\}_{i \geq 1}$ of breakpoints belonging to D and a sequence of numbers $\{\mu_i\}_{i \geq 1}$ such that

- a. $B_i \rightarrow B_0$ and $\mu_i \rightarrow 0$ as $i \rightarrow \infty$,*
- b. B_i minimizes $E(f_B, B)$ locally with $\mu = \mu_i$ and $\nu = \nu_0 \mu^2$,*
- c. $E(f_B, B) \rightarrow E_0(B)$ as $i \rightarrow \infty$.*

B. Conversely, suppose that B_0 does not locally minimize $E_0(B)$. Then there exists a constant μ_0 such that for all $\mu \leq \mu_0$, $E(f_B, B)$ does not achieve its global minimum at B_0 .

By analyzing the convergence of the derivatives of $E(f_B, B)$ and $E_0(B)$ as in [8], it is possible to obtain stronger statements.

5. Behavior when μ is large. By Theorem 1 again, in a nontrivial minimizing segmentation,

$$(\partial_R g)^2 \geq \frac{\nu}{2\mu} \quad \text{and} \quad (\text{Lip}_R g)^2 \geq \frac{\mu\nu}{2}.$$

Hence as $\mu \rightarrow \infty$, we must bound $\mu\nu$. We impose the condition

$$\frac{\mu\nu}{2} = \nu_\infty \leq (\text{Lip}_R g)^2$$

when μ is large. Let

$$E_\infty(B) = \sum_{x \in B} [\nu_\infty - \{g'(x)\}^2].$$

In [7] we show that for a fixed set B of breakpoints and $g \in C^{1,1}(R)$,

$$E(f_B, B) = E(f_\emptyset, \emptyset) + \frac{2}{\mu} E_\infty + O\left(\frac{1}{\mu^2}\right).$$

(This follows easily from Lemma 4 below.) This indicates that as $\mu \rightarrow \infty$, it becomes advantageous to place more and more breakpoints in the vicinity of points where $|g'(x)|$ is maximum. Note however that although $E_\infty(B)$ may have stationary points (where $g'' = 0$), unlike $E_0(B)$, it has no minima unless $\sqrt{\nu_\infty} > \|g'\|_{\infty, R}$. Thus we should expect $|B| \rightarrow \infty$ as $\mu \rightarrow \infty$.

Let $B_{\mu, \nu}$ denotes a set of breakpoints that minimizes $E(f_B, B)$ with fixed μ and ν .

THEOREM 4. *Suppose that $g \in C^{1,1}(R)$. Then, the following hold.*

i. For every $\varepsilon > 0$, there exists a constant $\mu_\varepsilon > 0$ such that for all $\mu > \mu_\varepsilon$ and $\nu = 2\nu_\infty/\mu$,

$$B_{\mu,\nu} \subset \{x: |g'(x)| \geq \sqrt{\nu_\infty} - \varepsilon\}.$$

ii. If $\|g'\|_{\infty,R} > \sqrt{\nu_\infty}$, then

$$|B_{\mu,\nu} \cap \{x: |g'(x)| > \sqrt{\nu_\infty}\}| \rightarrow \infty \quad \text{as } \mu \rightarrow \infty.$$

COROLLARY 3. Let $g \in C^{1,1}(R)$. Instead of assuming that $\mu\nu$ is fixed as $\mu \rightarrow \infty$, fix a positive integer k and suppose that for each μ we choose ν such that $|B_{\mu,\nu}| \leq k$. If $B_{\mu,\nu}$ is nonempty for all μ , then as $\mu \rightarrow \infty$,

$$\frac{\mu\nu}{2} \rightarrow \|g'\|_{\infty,R}^2 \quad \text{and} \quad \lim_{\mu \rightarrow \infty} B_{\mu,\nu} \subset \{x: |g'(x)| = \|g'\|_{\infty,R}\}.$$

Proof of Theorem 4. Let $c = \sqrt{\nu_\infty}$. Let $R_\varepsilon = \{x: |g'(x)| \leq c - \varepsilon/2\}$. Choose μ_0 such that

$$\left(c - \frac{\varepsilon}{2}\right)^2 < \frac{c^2}{1 - e^{-\mu_0|R|}}.$$

By Corollary 2, each connected component of R_ε contains at most two breakpoints of $B_{\mu,\nu}$ for all $\mu \geq \mu_0$. We show now that for sufficiently large μ , the breakpoints in R_ε can occur only at points where $|g'| \geq c - \varepsilon$. Consider a connected component W of R_ε which contains at least one point, P , of $B_{\mu,\nu}$ where $|g'| < c - \varepsilon$. Since W can have at most two points of $B_{\mu,\nu}$, it has an interval containing no breakpoints in which $c - \varepsilon \leq |g'(x)| \leq c - \varepsilon/2$. (This is true even if W contains an endpoint of R .) Let Q be a point in this interval such that $|g'(Q)| = c - 2\varepsilon/3$. We claim that if μ is sufficiently large, we can reduce the energy by removing the cut at P and placing it at Q . We need the following lemma, which is an extension of Lemma 1.

LEMMA 4. Let $D = [0, a]$ and let $g: D \rightarrow \mathbf{R}$ be a $C^{1,1}$ function. Let u satisfy the equation $u'' = \mu^2(u - g)$ in D and the boundary condition $u'(0) = u'(a) = 0$. Then there exists a constant C such that

$$\begin{aligned} \left|u(0) - g(0) - \frac{g'(0)}{\mu}\right| &\leq \frac{1}{\mu} \left[\left(\frac{1}{2} + \mu a\right) e^{-\mu a} \|g'\|_{\infty,D} + \frac{C}{\mu} (\text{Lip}_D g') \right], \\ |u'(x) - g'(x)| &\leq \left(\frac{3}{2} + \mu a\right) e^{-\mu a/2} \|g'\|_{\infty,D} + \frac{C}{\mu} (\text{Lip}_D g'). \end{aligned}$$

Proof of the Lemma. Extend g to all of \mathbf{R} by successive reflections of D . As in the proof of Lemma 1,

$$\begin{aligned} u(x) - g(x) &= \frac{\mu}{2} \int_{-\infty}^{\infty} e^{-\mu|y-x|} [g(y) - g(x)] dy \\ &= \mu e^{-\mu x} \int_0^x (\cosh \mu y) [g(y) - g(x)] dy \\ &\quad + \mu (\cosh \mu x) \int_x^{\infty} e^{-\mu y} [g(y) - g(x)] dy. \end{aligned}$$

Also from the proof of Lemma 1, we have

$$\begin{aligned} u'(x) &= -\mu^2 e^{-\mu x} \int_0^x (\cosh \mu y) [g(y) - g(x)] dy \\ &\quad + \mu^2 (\sinh \mu x) \int_x^{\infty} e^{-\mu y} [g(y) - g(x)] dy. \end{aligned}$$

The required estimates are now obtained by substituting

$$\begin{aligned} g(y) - g(x) &= g'(x)(y-x) + \tilde{g}(y)(y-x)^2 \quad \text{if } 0 \leq y \leq a, \\ |g(y) - g(x)| &\leq a \|g'\|_{\infty, D} \quad \text{if } y \geq a, \end{aligned}$$

where $|\tilde{g}(y)| \leq \text{Lip}_R g'$, and evaluating the integrals. \square

We continue now with the proof of Theorem 4. We may assume that the points P and Q are in the same segment S . Let $s = \varepsilon / (\text{Lip}_R g')$. Note that $|S| \geq s/2$. Choose $\mu_\varepsilon \geq \mu_0$ such that

$$\left(\frac{3}{2} + \frac{\mu_\varepsilon s}{12} \right) e^{-\mu_\varepsilon s/24} \|g'\|_{\infty, R} + \frac{C}{\mu_\varepsilon} (\text{Lip}_R g') \leq \frac{\varepsilon}{12},$$

where C is the constant defined in Lemma 4. Let S' be the other segment with P as one of its endpoints. Let f_0, f_1 , and f_2 be the solutions of the Neumann problem $f'' = \mu^2(f - g)$ in $S \cup S'$, $S \cup S' - \{P\}$, $S \cup S' - \{Q\}$, respectively, with homogeneous boundary conditions. Let $v_i = f_i - g$ for $i = 0, 1, 2$. Then the reduction in energy by moving the cut from P to Q equals

$$f'_0(Q)[v_2^+(Q) - v_2^-(Q)] - f'_0(P)[v_1^+(P) - v_1^-(P)].$$

By Lemma 4,

$$\begin{aligned} |f'_0(Q) - g'(Q)| &\leq \frac{\varepsilon}{12} \\ \left| v_2^\pm(Q) \mp \frac{g'(Q)}{\mu} \right| &\leq \frac{\varepsilon}{12\mu} \\ |f'_0(P)| &\leq |g'(P)| + \frac{\varepsilon}{12} \leq c - \frac{11\varepsilon}{12} \\ |v_1^+(P)| &\leq \frac{1}{\mu} \left(|g'(P)| + \frac{\varepsilon}{12} \right) \leq \frac{1}{\mu} \left(c - \frac{11\varepsilon}{12} \right). \end{aligned}$$

If $|S'| \geq s/2$, then

$$|v_1^-(P)| \leq \frac{1}{\mu} \left(c - \frac{11\varepsilon}{12} \right)$$

by Lemma 4. If $|S'| \leq s/2$, then

$$\|g'\|_{\infty, S'} \leq |g'(P)| + (\text{Lip}_R g') \frac{s}{12} \leq |g'(P)| + \frac{\varepsilon}{12}$$

and hence, by Lemma 1

$$|v_1^-(P)| \leq \frac{1}{\mu} \left(|g'(P)| + \frac{\varepsilon}{12} \right) \leq \frac{1}{\mu} \left(c - \frac{11\varepsilon}{12} \right)$$

again. Therefore

$$\text{Reduction in energy} \geq \frac{2}{\mu} \left(c - \frac{3\varepsilon}{4} \right)^2 - \frac{2}{\mu} \left(c - \frac{11\varepsilon}{12} \right)^2 > 0.$$

To prove part (ii) of the theorem, choose $\alpha > 0$ such that $N = \{x : c + \alpha < |g'(x)| < \|g'\|_{\infty, R}\}$ is not empty. Then, using Lemma 4, show in the same way as in part (i) that for every $\varepsilon > 0$, there exists μ_ε such that for all $\mu > \mu_\varepsilon$, if $I \subset N$ is an interval of length $\geq \varepsilon$, not containing a point of $B_{\mu, \nu}$, we can reduce the energy by placing a cut at the center of I . \square

REFERENCES

- [1] L. AMBROSIO, *Variational problems in SBV*, Center for Intelligent Control Systems, CICS-P-86, MIT, Cambridge, MA, 1988.
- [2] A. BLAKE AND A. ZISSERMAN, *Using weak continuity constraints*, Report CSR-186-85, Dept. of Comp. Sci, Edinburgh Univ., Edinburgh, UK, 1985.
- [3] G. DAL MASO, J. M. MOREL, AND S. SOLIMINI, *A variational method in image segmentation: Existence and approximation results*, CEREMADE-Universite Paris Dauphine, Paris, 1989, preprint.
- [4] E. DE GIORGI, M. CARRIERO, AND A. LEACI, *Existence theorem for a minimum problem with free discontinuity set*, Arch. Rational Mech. Anal., 108 (1989), pp. 195–218.
- [5] S. GEMAN AND D. GEMAN, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Trans. PAMI, 6 (1984) pp. 721–741.
- [6] J. L. MARROQUIN, *Surface reconstruction preserving discontinuities*, Artificial Intelligence Lab. Memo 792, MIT, Cambridge, MA, 1984.
- [7] D. MUMFORD AND J. SHAH, *Boundary Detection by Minimizing Functionals*, I, Proc. IEEE Conf. Computer Vision and Pattern Recognition, San Francisco, CA, 1985.
- [8] ———, *Optimal approximations by piecewise smooth functions and associated variational problems*, Comm. Pure Appl. Math., 42 (1989), pp. 577–684.
- [9] T. RICHARDSON, *Scale independent, piecewise smooth segmentation of images via variational methods*, Ph.D. thesis, Laboratory for Information and Decision Systems, LIDS-Th-1940, MIT, Cambridge, MA, 1990.

THE REVENGE OF THE LINEAR SEARCH PROBLEM*

ANATOLE BECK† AND MICAH BECK‡

Abstract. The linear search problem is the name for several problems motivated by the same external reality. At times, the search for a goal can proceed in two (or more) directions. Looking in one direction is at the expense of time and effort, which can be used elsewhere. More specifically, it might actually be moving further from the goal. This is modeled by a physical search along an infinite straight line, where the object of the search might be in either direction. Faced with a (known or unknown) probability distribution, this paper attempts to minimize the expected loss, where the loss is a function of the time of the search and the location of the object. In this variant of the problem, known distributions are dealt with, and the loss function is a known power of the time spent.

Key words. linear search, search games

AMS(MOS) subject classifications. 90B40, 90D45, 90D26, 93B40, 93C15, 93E20

Introduction. In 1972, the senior author, with the assistance of one of his students, took up the linear search problem under the assumption that the “cost” of each unit of distance increases with the time spent in search [4]. One way of defining the problem is to define the cost function $X_\alpha(x)$ as $\int_{-\infty}^{\infty} (X(x, t))^\alpha dF(t)$, where $x = \{x_i\}_{i=-\infty}^{\infty}$ is a search strategy, and $X(x, t)$ is defined as follows: for t lying between x_{n-1} and x_{n+1} , $X(x, t) = |t| + \sum_{i=-\infty}^n 2|x_i|$, and F is the (known or unknown) distribution of the target. Much has been written on the case where $\alpha = 1$ [1, 2, 3, 5, 6, 7]. Among the values of $\alpha > 1$, $\alpha = 2$ is special. This reflects not only the general mathematical interest in square-summable functions, but also the naive assumption that the function multiplying each new $dF(t)$ (possibly reflecting the degree of impatience) is proportional to the time already spent.

It will be the purpose of this paper to duplicate some of the results of [5] and [6] for the cases $\alpha > 1$ and especially $\alpha = 2$.

1. Definitions and fundamental notion. Let $\alpha > 1$ be arbitrarily fixed for the remainder of the paper and $\beta = \alpha - 1$. We will consider probability distributions F on the real line \mathbb{R} for which the absolute α -moment $M_\alpha = M_\alpha(F) = \int_{-\infty}^{\infty} |t|^\alpha dF(t) < \infty$. F is taken continuous from the left in \mathbb{R}^- , the negative reals, continuous from the right in \mathbb{R}^+ , and thus continuous at 0, for reasons set out in [1]. In [2], we define a *generalized search procedure* as a sequence $\{x_i\}_{i=-\infty}^{\infty}$ with

$$\cdots \leq x_2 \leq x_0 \leq x_{-2} \leq \cdots \leq 0 \leq \cdots \leq x_{-1} \leq x_1 \leq x_3 \leq \cdots.$$

We denote the set of generalized search plans as \mathcal{X}_1 . For each point $t \in \mathbb{R}$, if t lies between x_{n-1} and x_{n+1} , we consider that the search plan x envisions a path beginning at 0 and consisting of the intervals $\cdots, [x_{-2}, x_{-1}], [x_0, x_{-1}], [x_0, x_1], \cdots$, up to the point x_n , followed by the interval between x_n and t . The length of this path is $|t| + 2s_n$, where $s_n = \sum_{i=-\infty}^n |x_i|$. If $\sum_{i=-\infty}^0 |x_i| = \infty$, then $X(x, t) = \infty$, for all $t \in \mathbb{R}$, and x is a very unsatisfactory search procedure. Since the search procedure $y = \{-(-2)^n\}$ yields

* Received by the editors November 28, 1989; accepted for publication (in revised form) February 11, 1991.

† Department of Mathematics, University of Wisconsin, Madison, Wisconsin 53706. This research was supported by the Wisconsin Alumni Research Foundation and the London School of Economics.

‡ School of Computer and Information Science, Center for Science and Technology, Syracuse University, Syracuse, New York, 13244-4100. This research was supported by a Hewlett Packard Faculty Development Fellowship and by National Science Foundation grant DCR86-01864.

$X(x, t) < 9|t|$, for all $t \in \mathbb{R}$, we see that $X_\alpha(y) = \int_{-\infty}^{\infty} X(y, t)^\alpha dF(t) \leq 9^\alpha M_\alpha(F)$. Thus we have at least one strategy giving a finite value for X_α if $M_\alpha < \infty$. On the other hand, $X(x, t) \geq |t|$ for all search procedures x and all $t \in \mathbb{R}$, giving $X_\alpha(x) \geq M_\alpha(F)$ for all search procedures. By [4, Thm. 17], if $M_\alpha(F) < \infty$, then there is a search procedure z such that $X_\alpha(z) \leq X_\alpha(x)$ for all search procedures x . By [4, Thm. 29], if either $\bar{F}^+(0)$ or $\bar{F}^-(0)$ is finite, then there exists a $k \in \mathbb{Z}$ such that all the turning points z_i of z are 0 for $i \leq k$, where

$$\bar{F}^+(0) = \limsup \{(F(t) - F(0))/t \mid t \downarrow 0\}$$

and

$$\bar{F}^-(0) = \limsup \{(F(t) - F(0))/t \mid t \uparrow 0\}.$$

At this point, we will also prove the following lemma.

LEMMA 1.1. *If x is a minimizing search procedure for the distribution F and exponent $\alpha > 0$, then $|x_{n-1}| < |x_{n+1}|$ unless the search has not really begun ($x_n = x_{n+1} = 0$) or is really already over ($|F(x_n) - F(x_{n-1})| = 1$).*

Proof. Assume without loss of generality that $x_n \leq 0$; the other case is similar, mutatis mutandis. If $x_{n-1} = x_{n+1}$, let y be the search procedure

$$\begin{aligned} y_j &= x_j, & \forall j < n, \\ y_j &= x_{j+z}, & \forall j \geq n. \end{aligned}$$

Then $X(y, t) = X(x, t) - 2(|x_n| + |x_{n-1}|)$, $\forall t \notin (x_n, x_{n-1})$. Thus $X_\alpha(y) < X_\alpha(x)$ if $F(x_n) > 0$ or $F(x_{n-1}) < 1$. \square

2. Uniform distribution. We begin with the distribution F , defined by $F'(t) = 1/(b-a)$, for all $a < t < b$, $F'(t) = 0$ elsewhere. If $a > 0$ or $b < 0$, the problem is trivial, so we take $a < 0 < b$. To simplify the notation, we will rewrite the left-hand endpoint of (a, b) as $(-a, b)$ with $a > 0$, and rewrite each search strategy by omitting as turning points all x_i that are 0 or lie outside $[-a, b]$.

THEOREM 2.1. *If $a > b$, then $X_\alpha(x)$ is minimized by $x = \{b, -a\}$. If $a < b$, then $X_\alpha(x)$ is minimized by $\{-a, b\}$.*

We begin with the following lemma.

LEMMA 2.2. *If $0 < a < b$, then*

$$\int_0^a t^\alpha dt + \int_0^b (2a+t)^\alpha dt < \int_0^b t^\alpha dt + \int_0^a (2b+t)^\alpha dt.$$

Proof. We must show that

$$\int_0^b (2a+t)^\alpha - t^\alpha dt < \int_0^a (2b+t)^\alpha - t^\alpha dt,$$

i.e.,

$$\int_0^b \alpha \int_0^{2a} (t+s)^\beta ds dt < \int_0^a \alpha \int_0^{2b} (t+s)^\beta ds dt.$$

Ignoring the common factor α and subtracting the common domain $[0, a] \times [0, b]$, we must show that

$$\int_0^b \int_a^{2a} (t+s)^\beta ds dt < \int_0^a \int_b^{2b} (t+s)^\beta ds dt,$$

i.e.,

$$\begin{aligned} \int_0^b \int_0^a (a+t+s)^\beta ds dt &< \int_0^a \int_0^b (b+t+s)^\beta ds dt \\ &= \int_0^b \int_0^a (b+t+s)^\beta ds dt, \end{aligned}$$

which is clear since $a+t+s < b+t+s$, for all $s, t \in \mathbb{R}$. \square

LEMMA 2.3. *If $0 < a < b < c$, then*

$$\int_0^b t^\alpha dt + \int_0^c (2b+t)^\alpha dt < \int_0^a t^\alpha dt + \int_0^b (2a+t)^\alpha dt + \int_a^c (2a+2b+t)^\alpha dt.$$

Proof. We must show that

$$\int_0^a (2b+t)^\alpha - t^\alpha dt < \int_0^b (2a+t)^\alpha - t^\alpha dt + \int_a^c (2a+2b+t)^\alpha - (2b+t)^\alpha dt,$$

i.e.,

$$\begin{aligned} \int_0^a \int_0^{2b} (s+t)^\beta ds dt &< \int_0^b \int_0^{2a} (s+t)^\beta ds dt + \int_a^c \int_0^{2a} (2b+s+t)^\beta ds dt, \\ \int_a^c \int_0^a (2b+s+t)^\beta ds dt &> a(2b+a)^\beta (c-a) > a(2a+b)^\beta (b-a) > \int_{2a}^{a+b} \int_0^a (s+t)^\beta ds dt, \\ \int_a^c \int_a^{2a} (2b+s+t)^\beta ds dt &> a(2b+2a)^\beta (c-a) > a(2b+a)^\beta (b-a) > \int_{a+b}^{2b} \int_0^a (s+t)^\beta ds dt, \\ \int_0^b \int_0^{2a} (s+t)^\beta ds dt &> \int_0^{2a} \int_0^a (s+t)^\beta ds dt. \end{aligned}$$

Adding the extremes,

$$\begin{aligned} \int_a^c \int_0^{2a} (2b+s+t)^\beta ds dt + \int_0^b \int_0^{2a} (s+t)^\beta ds dt &> \int_0^{2b} \int_0^a (s+t)^\beta ds dt \\ &= \int_0^a \int_0^{2b} (s+t)^\beta ds dt. \end{aligned} \quad \square$$

LEMMA 2.4. *If $0 < b \leq a < c$, then the same conclusion holds as in Lemma 2.3.*

Proof. The inequality

$$\int_0^b t^\alpha dt + \int_0^a (2b+t)^\alpha dt \leq \int_0^a t^\alpha dt + \int_0^b (2a+t)^\alpha dt$$

holds by Lemma 2.2, while

$$\int_a^c (2b+t)^\alpha dt < \int_a^c (2a+2b+t)^\alpha dt. \quad \square$$

COROLLARY 2.5. *If $0 < a < c$ and $0 < b < c$, then the same conclusion holds.*

LEMMA 2.6. *If $0 < a < c \leq b < d$, then*

$$\begin{aligned} \int_0^c t^\alpha dt + \int_0^d (2c+t)^\alpha dt &< \int_0^a t^\alpha dt + \int_0^b (2a+t)^\alpha dt + \int_a^c (2a+2b+t)^\alpha dt \\ &\quad + \int_b^d (2a+2b+2c+t)^\alpha dt. \end{aligned}$$

Proof. Note that

$$\begin{aligned} \int_b^d (2a+2b+2c+t)^\alpha dt &> \int_b^d (2c+t)^\alpha dt, \\ \int_0^c t^\alpha dt - \int_0^a t^\alpha dt &= \int_a^c t^\alpha dt, \end{aligned}$$

and

$$\begin{aligned} \int_a^c (2a+2b+t)^\alpha - t^\alpha dt &= \alpha \int_a^c \int_0^{2a+2b} (s+t)^\beta ds dt \\ &> \alpha \int_a^c \int_{2a+b}^{2a+2b} (s+t)^\beta ds dt + \alpha \int_a^c \int_{2a}^{2a+b} (s+t)^\beta ds dt. \end{aligned}$$

On the other hand,

$$\begin{aligned} \int_0^b (2c+t)^\alpha - (2a+t)^\alpha dt &= \alpha \int_0^b \int_{2a}^{2c} (s+t)^\beta ds dt \\ &= \alpha \int_0^b \int_{2a}^{a+c} (s+t)^\beta ds dt + \alpha \int_0^b \int_{a+c}^{2c} (s+t)^\beta ds dt. \end{aligned}$$

Also,

$$\begin{aligned} \int_a^c \int_{2a+b}^{2a+2b} (s+t)^\beta ds dt &= \int_a^c \int_0^b (2a+b+s+t)^\beta ds dt \\ &> \int_a^c \int_0^b (c+s+t)^\beta ds dt \\ &= \int_0^b \int_{a+c}^{2c} (s+t)^\beta ds dt \end{aligned}$$

and

$$\begin{aligned} \int_a^c \int_{2a}^{2a+b} (s+t)^\beta ds dt &= \int_a^c \int_0^b (2a+s+t)^\beta ds dt \\ &> \int_a^c \int_0^b (a+s+t)^\beta ds dt \\ &= \int_0^b \int_{2a}^{a+c} (s+t)^\beta ds dt. \end{aligned}$$

Adding the first two inequalities gives us

$$\begin{aligned} \int_b^d (2a+2b+2c+t)^\alpha dt + \int_a^c (2a+2b+t)^\alpha dt - \int_a^c t^\alpha dt \\ > \int_b^d (2c+t)^\alpha dt + \alpha \int_a^c \int_{2a+b}^{2a+2b} (s+t)^\beta ds dt + \alpha \int_a^c \int_{2a}^{2a+b} (s+t)^\beta ds dt. \end{aligned}$$

Incorporating the first equality gives us

$$\begin{aligned} \int_b^d (2a+2b+2c+t)^\alpha dt + \int_a^c (2a+2b+t)^\alpha dt + \int_0^a t^\alpha dt \\ > \int_0^c t^\alpha dt + \int_b^d (2c+t)^\alpha dt + \alpha \int_a^c \int_{2a+b}^{2a+2b} (s+t)^\beta ds dt + \alpha \int_a^c \int_{2a}^{2a+b} (s+t)^\beta ds dt. \end{aligned}$$

It remains to show that

$$\int_0^b (2a+t)^\alpha dt > \int_0^b (2c+t)^\alpha dt - \alpha \int_a^c \int_{2a+b}^{2a+2b} (s+t)^\beta ds dt - \alpha \int_a^c \int_{2a}^{2a+b} (s+t)^\beta ds dt,$$

which, by the second equality, means

$$\begin{aligned} \alpha \int_a^c \int_{2a+b}^{2a+2b} (s+t)^\beta ds dt + \alpha \int_a^c \int_{2a}^{2a+b} (s+t) ds dt \\ > \alpha \int_0^b \int_{2a}^{a+c} (s+t)^\beta ds dt + \alpha \int_0^b \int_{a+c}^{2c} (s+t)^\beta ds dt, \end{aligned}$$

which follows from the last two inequalities. \square

Proof of Theorem 2.1. By [4, Thm. 29], there is a minimizing strategy $\{x_i\}$ with $x_i = 0$, for all $i \leq 0$. We cannot have $0 \neq x_1 \in (-a, b)$ and $|x_2| < |x_3|$, since in that case, by either Lemma 2.3 (if $|x_1| < |x_2|$) or Lemma 2.4 (if $|x_1| \geq |x_2|$), the substitution of 0 for x_1 would reduce $X_\alpha(x)$. If $|x_2| > |x_3|$ and $x_4 \in (-a, b)$, then by Lemma 2.6, the substitution of 0 for x_1 and x_2 would reduce $X_\alpha(x)$. Finally, if $|x_2| > |x_3|$ and $x_4 \notin (-a, b)$, the substitution of 0 for x_1 and x_2 , and of the endpoint (either $-a$ or b) of the same sign for x_4 would reduce $X_\alpha(x)$. Thus, in all cases, $X_\alpha(x)$ is not minimized if $0 \neq x_1 \in (-a, b)$. It follows that the first nonzero turning point is one of the endpoints, from which we see at once that the second is the other endpoint. By Lemma 2.2, x_1 must be the endpoint with the smaller absolute value. \square

3. Symmetric distributions. A distribution F is called *symmetric* if $F(x) + F(-x) = 1$, for all $x \in \mathbb{R}$. For these distributions, we take the notational convenience of writing our search strategies with positive entries (i.e., $\{|x_i|\}$ for $\{x_i\}$) with the understanding that the actual turning points alternate in sign. Clearly, x and $-x$ give the same value of X_α .

THEOREM 3.1. *If F is a symmetric distribution, x is a minimizing search strategy, and x_k satisfies $0 < x_k$ and $F(x_k) < 1$, then $x_{k+1} > x_k$.*

Proof. The proof of this theorem is a modification of the proof of Lemma 3.2.

LEMMA 3.2. *On the hypothesis of Theorem 3.1, $x_{k+1} \neq x_k$.*

Proof. Assume that $x_k = x_{k+1}$. Then when the search reaches the $(k+1)$ th turning point, the strategy x calls for crossing back to x_k before searching new territory. However, because of symmetry, the same search effectively can be accomplished without crossing back. More explicitly, define the strategy y by

$$\begin{aligned} y_n &= x_n, & \forall n \leq k, \\ y_n &= x_{n+1}, & \forall n > k. \end{aligned}$$

Note that all turning points after the k th are in the opposite half of \mathbb{R} . Indeed, we see that $X(x, t) = X(y, t)$, for all $|t| \leq x_k$, $X(x, t) = X(y, -t) + 2|x_k|$, for all $|t| > x_k$. Thus $X_\alpha(y) < X_\alpha(x)$, contrary to the minimality of x . \square

Proof of Theorem 3.1. In the proof of Lemma 3.2, we drop the entry x_{k+1} . In this proof, we assume that $x_{k+1} \leq x_k$, which means $x_{k+1} < x_k$. Let j be the largest value of n for which $x_n \leq x_k$. Thus $j \geq k+1$ and is of opposite parity to k . Then $x_k < x_{j+2}$, and we drop all the x_n for $k < n \leq j$. Thus, as before, if $y = (\dots, x_{k-1}, x_k, x_{j+1}, x_{j+2}, \dots)$, all the entries after the k th lie in the opposite half of \mathbb{R} . Then we have

$$\begin{aligned} X_\alpha(x) &= \sum_{n=-\infty}^{k-1} \int_{x_{n-1}}^{x_{n+1}} (2s_n + t)^\alpha dF(t) + \int_{x_{k-1}}^{x_{k+1}} (2s_k + t)^\alpha dF(t) \\ &\quad + \dots + \int_{x_j}^{x_{j+2}} (2s_{j+1} + t)^\alpha dF(t) + \sum_{n=j+2}^{\infty} \int_{x_{n-1}}^{x_{n+1}} (2s_n + t)^\alpha dF(t), \\ X_\alpha(y) &= \sum_{n=-\infty}^{k-1} \int_{x_{n-1}}^{x_{n+1}} (2s_n + t)^\alpha dF(t) + \int_{x_{k-1}}^{x_{j+1}} (2s_k + t)^\alpha dF(t) + \int_{x_k}^{x_{j+2}} (2r_{j+1} + t)^\alpha dF(t) \\ &\quad + \sum_{n=j+2}^{\infty} \int_{x_{n-1}}^{x_{n+1}} (2r_n + t)^\alpha dF(t), \end{aligned}$$

where

$$s_n = \sum_{i=-\infty}^n x_i, \quad \forall n \in \mathbb{Z}, \quad \text{and} \quad r_n = s_k + x_{j+1} + \cdots + x_n < s_n, \quad \forall n > j.$$

Then

$$\begin{aligned} X_\alpha(x) - X_\alpha(y) &= \int_{x_{k-1}}^{x_{k+1}} (2s_k + t)^\alpha dF(t) + \int_{x_{k+1}}^{x_{k+3}} (2s_{k+2} + t)^\alpha dF(t) \\ &\quad + \cdots + \int_{x_{j-2}}^{x_j} (2s_{j-1} + t)^\alpha dF(t) + \int_{x_j}^{x_k} (2s_{j+1} + t)^\alpha dF(t) \\ &\quad + \int_{x_k}^{x_{j+2}} (2s_{j+1} + t)^\alpha dF(t) + \int_{x_k}^{x_{k+2}} (2s_{k+1} + t)^\alpha dF(t) \\ &\quad + \int_{x_{k+2}}^{x_{k+4}} (2s_{k+3} + t)^\alpha dF(t) + \cdots + \int_{x_{j-1}}^{x_{j+1}} (2s_j + t)^\alpha dF(t) \\ &\quad - \int_{x_{k-1}}^{x_k} (2s_k + t)^\alpha dF(t) - \int_{x_k}^{x_{j+1}} (2s_k + t)^\alpha dF(t) \\ &\quad - \int_{x_k}^{x_{j+2}} (2r_{j+1} + t)^\alpha dF(t) + \sum_{n=j+2}^{\infty} \int_{x_{n-1}}^{x_{n+1}} (2s_n + t)^\alpha - (2r_n + t)^\alpha dF(t) \\ &> \int_{x_{k-1}}^{x_{k+1}} (2s_k + t)^\alpha dF(t) + \int_{x_{k+1}}^{x_{k+3}} (2s_k + t)^\alpha dF(t) \\ &\quad + \cdots + \int_{x_{j-2}}^{x_j} (2s_k + t)^\alpha dF(t) + \int_{x_j}^{x_k} (2s_k + t)^\alpha dF(t) \\ &\quad + \int_{x_k}^{x_{j+2}} (2s_{j+1} + t)^\alpha dF(t) + \int_{x_k}^{x_{k+2}} (2s_k + t)^\alpha dF(t) \\ &\quad + \cdots + \int_{x_{j-1}}^{x_{j+1}} (2s_k + t)^\alpha dF(t) - \int_{x_{k-1}}^{x_{k+1}} (2s_k + t)^\alpha dF(t) \\ &\quad - \int_{x_{k+1}}^{x_k} (2s_k + t)^\alpha dF(t) - \int_{x_k}^{x_{j+1}} (2s_k + t)^\alpha dF(t) \\ &\quad - \int_{x_k}^{x_{j+2}} (2r_{j+1} + t)^\alpha dF(t) \\ &= \int_{x_{k+1}}^{x_k} (2s_k + t)^\alpha dF(t) + \int_{x_k}^{x_{j+2}} (2s_{j+1} + t)^\alpha dF(t) \\ &\quad + \int_{x_k}^{x_{j+1}} (2s_k + t)^\alpha dF(t) - \int_{x_{k+1}}^{x_k} (2s_k + t)^\alpha dF(t) \\ &\quad - \int_{x_k}^{x_{j+2}} (2r_{j+1} + t)^\alpha dF(t) - \int_{x_k}^{x_{j+1}} (2s_k + t)^\alpha dF(t) \\ &= \int_{x_k}^{x_{j+2}} (2s_{j+1} + t)^\alpha - (2r_{j+1} + t)^\alpha dF(t) > 0, \end{aligned}$$

contrary to the assumed minimality of $X_\alpha(x)$. \square

4. The triangular distribution. We define the triangular distribution T by the condition that $\dot{T}(t) = 1 - |t|$, for all $-1 < t < 1$, $\dot{T}(t) = 0$, for all $|t| > 1$. The only result in this section is Theorem 4.1.

THEOREM 4.1. *If x is a minimizing search strategy for T , then $x_i < 1$, for all $i \in \mathbb{Z}$.*

Proof. Suppose that k is the least value of n for which $x_n = 1$, i.e., $x = \{x_1, \dots, x_{k-1}, 1, 1, \dots\}$. Let $y = \{x_1, \dots, x_{k-1}, s, 1, 1, \dots\}$. Then

$$\begin{aligned} X_\alpha(x) = \sum_{n=0}^{k-2} \int_{x_{n-1}}^{x_{n+1}} (2s_n + t)^\alpha (1-t) dt + \int_{x_{k-2}}^1 (2s_{k-1} + t)^\alpha (1-t) dt \\ + \int_{x_{k-1}}^1 (2s_{k-1} + 2 + t)^\alpha (1-t) dt \end{aligned}$$

and

$$\begin{aligned} X_\alpha(y) = \sum_{n=0}^{k-2} \int_{x_{n-1}}^{x_{n+1}} (2s_n + t)^\alpha (1-t) dt + \int_{x_{k-2}}^s (2s_{k-1} + t)^\alpha (1-t) dt \\ + \int_{x_{k-1}}^1 (2s_{k-1} + 2s + t)^\alpha (1-t) dt \\ + \int_s^1 (2s_{k-1} + 2s + 2 + t)^\alpha (1-t) dt. \end{aligned}$$

Thus

$$\begin{aligned} X_\alpha(x) - X_\alpha(y) &= \int_s^1 (2s_{k-1} + t)^\alpha (1-t) dt \\ &+ \int_{x_{k-1}}^1 ((2s_{k-1} + 2 + t)^\alpha - (2s_{k-1} + 2s + t)^\alpha) (1-t) dt - \int_s^1 (2s_{k-1} + 2s + 2 + t)^\alpha (1-t) dt \\ &= \int_{x_{k-1}}^1 ((2s_{k-1} + 2 + t)^\alpha - (2s_{k-1} + 2s + t)^\alpha) (1-t) dt \\ &- \int_s^1 ((2s_{k-1} + 2s + 2 + t)^\alpha - (2s_{k-1} + t)^\alpha) (1-t) dt. \end{aligned}$$

By the mean value theorem,

$$(2s_{k-1} + 2 + t)^\alpha - (2s_{k-1} + 2s + t)^\alpha \geq 2\alpha(1-s)(2s_{k-1} + 2s)^\beta > 2\alpha(1-s)(2s_{k-1})^\beta,$$

while

$$(2s_{k-1} + 2s + 2 + t)^\alpha - (2s_{k-1} + t)^\alpha < 4\alpha(1+s)(2s_{k-1} + 2s + 3)^\beta < 8\alpha(2s_{k-1} + 5)^\beta.$$

Hence,

$$X_\alpha(x) - X_\alpha(y) > \alpha(1-s)(2s_{k-1})^\beta(1-x_{k-1})^2 - 4\alpha(2s_{k-1} + 5)^\beta(1-s)^2 > 0,$$

if $1-s$ is sufficiently small, contrary to the assumed minimality of $X_\alpha(x)$. \square

5. The special case $\alpha = 2$. For a symmetric distribution F , define the function H in \mathbb{R}^+ by $H(s) = \int_0^s t dF(t)$. Then for any search plan $x = \{x_n\}_{n=-\infty}^\infty$, we have

$$X_2(x) = \sum_{n=-\infty}^\infty \int_{x_{n-1}}^{x_{n+1}} (2s_n + t)^2 dF(t) = M_2(F) + 4\Delta_1 + 4\Delta_2,$$

where

$$\begin{aligned}
 M_2(F) &= \int_{-\infty}^{\infty} t^2 dF(t) = \sum_{n=-\infty}^{\infty} \int_{x_{n-1}}^{x_{n+1}} t^2 dF(t), \\
 \Delta_1 &= \sum_{n=-\infty}^{\infty} \int_{x_{n-1}}^{x_{n+1}} s_n t dF(t) \\
 &= \sum_{n=-\infty}^{\infty} s_n (H(x_{n+1}) - H(x_{n-1})) \\
 &= \sum_{n=-\infty}^{\infty} \sum_{j=-\infty}^n x_j (H(x_{n+1}) - H(x_{n-1})) \\
 &= \sum_{-\infty < j \leq n} x_j (H(x_{n+1}) - H(x_{n-1})) \\
 &= \sum_{j=-\infty}^{\infty} x_j (2H_{\infty} - H(x_{j-1}) - H(x_j)),
 \end{aligned}$$

where

$$H_{\infty} = \int_0^{\infty} t dF(t) = \frac{1}{2} M_1(F),$$

and

$$\Delta_2 = \sum_{n=-\infty}^{\infty} \int_{x_{n-1}}^{x_{n+1}} s_n^2 dF(t) = \sum_{n=-\infty}^{\infty} s_n^2 (F(x_{n+1}) - F(x_{n-1})).$$

If F is differentiable, then a minimizing search strategy x would satisfy $\partial \Delta / \partial x_k = 0$, for all $k \in \mathbb{Z}$, where $\Delta = \Delta_1 + \Delta_2$. The corresponding formula in the case where $\alpha = 1$ involves only the three turning points x_{k-1} , x_k , and x_{k+1} , thus enabling it to be used easily as a computing tool. For $\alpha = 2$, however, the matter is more complicated; see below:

$$\begin{aligned}
 \frac{\partial \Delta_1}{\partial x_k} &= (2H_{\infty} - H(x_{k-1}) - H(x_k)) - H'(x_k)x_k - H'(x_k)x_{k+1} \\
 &= 2H_{\infty} - H(x_{k-1}) - H(x_k) - x_k F'(x_k)(x_k + x_{k+1}), \\
 \frac{\partial \Delta_2}{\partial x_k} &= \sum_{n=k}^{\infty} 2s_n (F(x_{n+1}) - F(x_{n-1})) + F'(x_k)s_{k-1}^2 - F'(x_k)s_{k+1}^2 \\
 &= \sum_{n=k}^{\infty} 2s_k (F(x_{n+1}) - F(x_{n-1})) + \sum_{n=k}^{\infty} 2(s_n - s_k)(F(x_{n+1}) - F(x_{n-1})) - F'(x_k)(s_{k+1}^2 - s_{k-1}^2) \\
 &= 2s_k (G(x_{k-1}) + G(x_k)) + \sum_{n=k}^{\infty} 2 \sum_{j=k+1}^n x_j (F(x_{n+1}) - F(x_{n-1})) \\
 &\quad - F'(x_k)(s_{k+1} + s_{k-1})(s_{k+1} - s_{k-1}) \\
 &= 2s_k (G(x_{k-1}) + G(x_k)) + 2 \sum_{k < j \leq n} x_j (F(x_{n+1}) - F(x_{n-1})) \\
 &\quad - F'(x_k)(s_{k+1} + s_{k-1})(x_{k+1} + x_k) \\
 &= 2s_k (G(x_{k-1}) + G(x_k)) + 2 \sum_{j=k+1}^{\infty} x_j (G(x_{j-1}) + G(x_j)) - F'(x_k)(s_{k+1} + s_{k-1})(x_{k+1} + x_k),
 \end{aligned}$$

where

$$G(x) = 1 - F(x), \quad \forall x > 0.$$

Thus, our necessary condition becomes

$$\begin{aligned} 0 = \frac{\partial \Delta}{\partial x_k} &= 2H_\infty - H(x_{k-1}) - H(x_k) - x_k F'(x_k)(x_k + x_{k+1}) - (s_{k+1} + s_{k-1}) F'(x_k)(x_k + x_{k+1}) \\ &\quad + 2s_k(G(x_{k-1}) + G(x_k)) + 2 \sum_{j=k+1}^{\infty} x_j(G(x_{j-1}) + G(x_j)) \\ &= 2H_\infty - H(x_{k-1}) - H(x_k) - (s_{k+1} + s_k) F'(x_k)(x_{k+1} + x_k) + 2s_k(G(x_{k-1}) + G(x_k)) \\ &\quad + 2 \sum_{j=k+1}^{\infty} x_j(G(x_{j-1}) + G(x_j)). \end{aligned}$$

Since this is true for all $k \in \mathbb{Z}$,

$$\begin{aligned} 0 = \frac{\partial \Delta}{\partial x_k} - \frac{\partial \Delta}{\partial x_{k+1}} &= H(x_{k+1}) - H(x_{k-1}) - (s_{k+1} + s_k) F'(x_k)(x_{k+1} + x_k) \\ &\quad + (s_{k+2} + s_{k+1}) F'(x_{k+1})(x_{k+2} + x_{k+1}) + 2s_k(G(x_{k-1}) + G(x_k)) \\ &\quad - 2s_{k+1}(G(x_k) + G(x_{k+1})) + 2x_{k+1}(G(x_k) + G(x_{k+1})) \\ &= H(x_{k+1}) - H(x_{k-1}) - (s_{k+1} + s_k) F'(x_k)(x_{k+1} + x_k) \\ &\quad + (s_{k+2} + s_{k+1}) F'(x_{k+1})(x_{k+2} + x_{k+1}) + 2s_k(G(x_{k-1}) - G(x_{k+1})), \end{aligned}$$

which has the advantage that it involves only finitely many values of x_n . We can, in principle, use this formula to calculate x_{k+2} when x_1, \dots, x_{k+1} are known. Actually, the calculation is very delicate, depending on the size of $F'(x_{k+1})$. However, for every choice for x_1 and x_2 , we can generate values for all the other x_n . Thus, what would be a one-dimensional search in the case where $\alpha = 1$ becomes a two-dimensional search for $\alpha = 2$. We will carry out a numerical analysis for the triangular and normal distributions.

6. Numerical approximations. To obtain the turning points for the triangular distribution, we start with the search strategy $(1, 1, 1, 1, \dots)$. In line with Theorem 4.1, we modify $x_1 = 1$ to x_{11} , which satisfies the equation for $\partial \Delta / \partial x_1 = 0$, with $x_0 = 0$ and $x_2 = x_3 = 1$. Then we modify the strategy $(x_{11}, 1, 1, 1, \dots)$ by applying the equation for $\partial \Delta / \partial x_2 = 0$, with $x_1 = x_{11}$ and $x_3 = x_4 = 1$. The solution is called x_{22} , and we then modify x_{11} to x_{21} , using $x_0 = 0$, $x_2 = x_{22}$, and $x_3 = x_4 = 1$. We continue in this way, generating $x_{33}, x_{32}, x_{31}, x_{44}$, etc.

Actually, solving for x_{kj} will give us $x_{kj} = 1$ for quite small values of k and j because of the limitations of the machine, so we alter our algorithm to yield $\delta_{kj} = 1 - x_{kj}$ instead. Then we obtain $\delta_n = \lim_{k \rightarrow \infty} \delta_{kn}$, for all $n \in \mathbb{N}$. This procedure will give us the values shown in Table 1. For $\alpha = 1$,

$$X_1(x) = M_1(T) + \sum_{n=1}^{\infty} s_n((1 - x_{n-1})^2 - (1 - x_{n+1})^2) = M_1(T) + \sum_{n=0}^{\infty} (x_n + x_{n+1})(1 - x_n)^2.$$

Thus, the condition $0 = \partial X_1 / \partial x_n = (1 - x_n)^2 - 2(x_n + x_{n+1})(1 - x_n) + (1 - x_{n-1})^2$ translates as $0 = 3\delta_n^2 - (4 - 2\delta_{n+1}) + \delta_{n-1}^2$. We then seek a sequence $\{\delta_n\}$ satisfying this equation, for all $n \in \mathbb{N}$, and $\delta_0 = 1 - x_0 = 1$. For other values of α , we have

$$\frac{\partial X_\alpha}{\partial x_n} = \sum_{j=n}^{\infty} \int_{x_{j-1}}^{x_{j+1}} (2s_j + t)^\alpha (1 - t) dt - \delta_n((2s_{n+1} + x_n)^\alpha - (2s_{n-1} + x_n)^\alpha),$$

TABLE 1

α	$10\delta_1$	$100\delta_2$	$10^4\delta_3$	$10^8\delta_4$	$10^{17}\delta_5$	$10^{34}\delta_6$	$10^{68}\delta_7$	$10^{137}\delta_8$	$10^{276}\delta_9$
1.0	3.439070247	3.025814017	2.289280691	1.310201534	4.291570147	4.604393581	5.300110062	70.22791667	123299.0070
1.1	3.469181765	3.087519811	2.387234888	1.426001939	5.086643270	6.471094135	10.47187862	274.2132197	1880162.304
1.2	3.496987676	3.144237738	2.478693277	1.538434367	5.923038217	8.776889787	19.26863924	928.5782762	21563377.02
1.3	3.522091458	3.194920020	2.561429945	1.643672899	6.763277649	11.44625924	32.77697387	2687.258079	180610185.6
1.4	3.544193586	3.238712600	2.633485226	1.737970629	7.563020445	14.31522932	51.27215193	6576.093045	1081647293.
1.5	3.563076698	3.274944601	2.693208449	1.817874425	8.274965463	17.13798399	73.48841007	13509.94770	4565264038.
1.6	3.578591843	3.303115735	2.739287956	1.880423339	8.853637071	19.61784575	96.29143999	23194.24966	13455830642
1.7	3.590645396	3.322880994	2.770765467	1.923309790	9.260298286	21.45849747	115.1968310	33193.48990	27556843346
1.8	3.599186427	3.34032551	2.787034002	1.944988693	9.467183812	22.42298510	125.7641476	39557.63673	39132769450
1.9	3.604194377	3.36479163	2.787820234	1.944727148	9.460376733	22.38371584	125.2949899	39256.06376	38532943959
2.0	3.605666897	3.330223518	2.773153371	1.922594932	9.240928179	21.34868840	113.9416240	32456.73412	26335989506
2.1	3.603607679	3.315338000	2.743323447	1.879402602	8.824160139	19.45686156	94.60795965	22370.41556	12508119051
2.2	3.598014021	3.291939304	2.698832294	1.816598866	8.237410065	16.94552152	71.73078778	12855.43225	4129560077.
2.3	3.588863860	3.260162266	2.640340644	1.736141705	7.516703040	14.10060167	49.64280401	6154.929142	946336822.4
2.4	3.576102004	3.220133393	2.568614966	1.640358738	6.702938377	11.20429719	31.32617571	2449.844177	149874762.1
2.5	3.559625505	3.171944961	2.484478192	1.531812027	5.838154888	8.492567814	17.98644479	807.2462237	16266661.37
2.6	3.539268694	3.115631622	2.388769692	1.413181439	4.962321052	6.129906101	9.364321181	218.6946594	1193384.729
2.7	3.514789642	3.051153648	2.282322330	1.287179232	4.110923140	4.202640103	4.398320885	48.21790278	57985.78405
2.8	3.485862263	2.978394716	2.165968159	1.156506201	3.313433867	2.727237874	1.850696343	8.531642516	1814.488258
2.9	3.452082077	2.897187544	2.040588115	1.023854406	2.592564354	1.667679084	.6914070473	1.189972312	35.28010487
3.0	3.412998093	2.807385305	1.907220402	0.891948748	1.964058620	.9559032418	.2269498425	.1281195272	.4087310260

TABLE 2

α	x_1	x_2	x_3	x_4	x_5
1	1.44085411	2.62758012	3.63220012	4.52034243	5.32668134
2	1.73460565	2.91667028	3.90190853	4.77195919	5.56291993
	x_6	x_7	x_8	x_9	x_{10}
1	6.07157768	6.76811167	7.42526253	8.04950858	8.64570659
2	6.29484181	6.98033174	7.62796854	8.24392182	8.83281415
	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
1	9.21761012	9.76819253	10.2998576	10.8145824	11.3140164
2	9.39822060	9.94297686	10.4693791	10.9793191	11.4743781
	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}
1	11.7995539	12.2723858	12.7335393	13.1839076	13.6241273
2	11.9558947	12.4250141	12.8827259	13.3298920	13.7672691

which we wish to equal 0. Thus, we want

$$\begin{aligned}
 0 &= \frac{\partial X_\alpha}{\partial x_n} - \frac{\partial X_\alpha}{\partial x_{n+1}} = \int_{x_{n-1}}^{x_{n+1}} (2s_n + t)^\alpha (1-t) dt - \delta_n ((2s_{n+1} + x_n)^\alpha - (2s_{n-1} + x_n)^\alpha) \\
 &\quad + \delta_{n+1} ((2s_{n+2} + x_{n+1})^\alpha - (2s_n + x_{n+1})^\alpha) \\
 &= \frac{2}{\alpha+1} ((2s_n + x_{n+1})^\alpha (1 + \alpha\delta_{n+1} + 2s_n) \\
 &\quad - (2s_n + x_{n-1})^\alpha (1 + \alpha\delta_{n-1} + 2s_n)) \\
 &\quad - \delta_n ((2s_{n+1} + x_n)^\alpha - (2s_{n-1} + x_n)^\alpha) \\
 &\quad + \delta_{n+1} ((2s_{n+2} + x_{n+1})^\alpha - (2s_n + x_{n+1})^\alpha),
 \end{aligned}$$

where $x_n = 1 - \delta_n$ and $s_n = x_1 + \dots + x_n$, $\forall n \in \mathbb{N}$.

We obtain δ_n for each value of α between 1 and 3 by taking as initial values what we have for $\alpha = 1$ and applying Newton's method until we get the necessary zeroes. The results are found in Table 1.

The normal distribution is more complicated. To avoid the complications of infinitely many turning points, we will use the equations we get from the condition $(\partial\Delta/\partial x_k) - (\partial\Delta/\partial x_{k+1}) = 0$, for all $k \in \mathbb{N}$. Choosing values of x_1 and x_2 with $0 < x_1 < x_2$, we generate x_3, \dots, x_{20} recursively and calculate Δ for those x_1, x_2 for which $0 < x_1 < x_2 < \dots < x_{20} < 10^3$. Table 2 shows a comparison of the turning points for $\alpha = 1, 2$.

REFERENCES

- [1] A. BECK, *On the linear search problem*, Israel J. Math., 2 (1964), pp. 221–228.
- [2] ———, *More on the linear search problem*, Israel J. Math., 3 (1965), pp. 61–70.
- [3] A. BECK AND D. J. NEWMAN, *Yet more on the linear search problem*, Israel J. Math., 8 (1970), pp. 419–429.
- [4] A. BECK AND P. WARREN, *The return of the linear search problem*, Israel J. Math., 14 (1973), pp. 503–512.
- [5] A. BECK AND M. BECK, *Son of the linear search problem*, Israel J. Math., 48 (1984), pp. 109–122.
- [6] ———, *The linear search problem rides again*, Israel J. Math., 53 (1986), pp. 365–372.
- [7] P. J. ROUSSEUW, *Optimal search paths for random variables*, J. Comput. Appl. Math., 9 (1983), pp. 279–286.

H_∞ -CONTROL BY STATE-FEEDBACK FOR PLANTS WITH ZEROS ON THE IMAGINARY AXIS*

CARSTEN SCHERER†

Abstract. Algebraic tests are derived for the suboptimality of some parameter in the H_∞ -optimization problem by state-feedback, where the finite zero structure of the plant is not restricted. As an application of these characterizations, a quadratically convergent algorithm for the computation of the optimal value is presented. The suboptimality tests are based on new solvability criteria for general algebraic Riccati inequalities that are of independent interest.

Key words. H_∞ -optimization, invariant zeros, Riccati inequalities, quadratic convergence

AMS(MOS) subject classifications. 93C05, 93C35, 93C45, 93C60, 93D15, 49B99

1. Introduction. In the seminal papers [15], [10], [11], and [28], suboptimality tests for the state-feedback H_∞ -problem were derived by making essential use of the bounded real lemma [2]. These characterizations are formulated in terms of the solvability of an algebraic Riccati *equation* (ARE) that contains some perturbation parameter $\varepsilon > 0$. For the so-called regular problem (in the terminology of [23]), this ε -parametrized Riccati equation could be viewed as an unperturbed strict algebraic Riccati *inequality* (ARI). If the plant has, in addition, no zeros on the imaginary axis, we can return to an unperturbed Riccati *equation* [3], [24] whose solvability can be checked algebraically. The ARE-based suboptimality tests allow us to find quadratically convergent algorithms for computing the optimal value [18]. If considering the *singular* problem for a plant without zeros on the imaginary axis, the perturbation technique can be avoided as well. We must replace the Riccati equation by a quadratic matrix inequality as it is known in the theory of singular LQ problems [23]. In fact, the suboptimality criteria boil down to the solvability of a certain reduced-order Riccati equation, which implies that the computational algorithms of [18] are again applicable.

In this paper we provide an algebraic solvability test for general strict Riccati inequalities. This leads directly to a characterization of suboptimality for the regular H_∞ -problem without the need to apply any perturbation technique. In a subsequent paper we show how these results may be generalized to a solution of the H_∞ -problem by output measurement [19].

The paper is structured as follows. In § 2 we give a precise formulation of the regular H_∞ -optimization problem by state-feedback and point out the relation of suboptimality with algebraic Riccati inequalities (Theorem 1). We briefly explain the above-described perturbation methods and the difficulties caused by plant zeros on the imaginary axis. The self-contained § 3 is the core of our whole approach. It contains the derivation of new algebraic conditions which are equivalent to the solvability of a strict algebraic Riccati inequality with a symmetric or positive definite matrix (Theorems 3 and 6). The parameter matrices that define the ARI are in no way restricted. Section 4 consists of the translation of these criteria to the H_∞ -problem. Although this is just a task of matching matrices, we explicitly discuss the influence of the plant zeros on the H_∞ -optimal value and on the design of suboptimal feedbacks. Section 5

*Received by the editors January 5, 1990; accepted for publication (in revised form) December 13, 1990.

†Mathematisches Institut, Am Hubland, D-8700 Würzburg, Germany. This author was supported by Deutscher Akademischer Austauschdienst. This work was conducted while the author was visiting the Mathematics Institute of the University of Groningen, the Netherlands.

is devoted to proposing an algorithm for the computation of the optimal value. Moreover, we investigate the possibility of computing the optimal value by solving certain Hermitian eigenvalue problems.

1b. Notation. We denote by \mathbb{N} the nonnegative integers, by \mathbb{R} and \mathbb{C} the real and complex numbers where \mathbb{C} is partitioned in the usual way as $\mathbb{C}^- \cup \mathbb{C}^0 \cup \mathbb{C}^+$, the open left half-plane, the imaginary axis and the open right half-plane, respectively. The spaces $\mathbb{R}^n, \mathbb{C}^n$ are equipped with the usual Euclidean norm and $\mathbb{R}^{n \times m}, \mathbb{C}^{n \times m}$ with the induced operator norm where all these norms are denoted by $\|\cdot\|$. Any matrix and any subspace appearing in this paper is considered to be *real* if not stated otherwise. In general, the dimensions of (sub)matrices are suppressed and a block in a partitioned matrix which is of no interest is denoted as “*.” Moreover, we use the notation $A|_{\mathcal{J}}$ for the restriction of $A \in \mathbb{R}^{n \times n}$ to any A -invariant subspace \mathcal{J} .

For the subset of symmetric matrices in $\mathbb{R}^{n \times n}$ we use the symbol \mathbb{S}^n . $X > Y$ ($X \geq Y$) means that X and Y are symmetric and $X - Y$ is positive (semi)definite. If \mathcal{S} is any subset of \mathbb{S}^n , S_- is called a (strict) lower bound if $S_- \leq S$ ($S_- < S$) holds for all $S \in \mathcal{S}$. If there exists one (strict) lower bound, there are obviously infinitely many. An important concept is to choose “close” (strict) lower bounds: S_- is called a (strict) lower limit point of \mathcal{S} if S_- is a (strict) lower bound of \mathcal{S} and if there exists a sequence $S_j \in \mathbb{S}^n$ with $S_j \rightarrow S_-$ for $j \rightarrow \infty$. Obviously, there is *at most one* lower limit point and *at most one* strict lower limit point of \mathcal{S} . A matrix $S \in \mathbb{S}^n$ is called positive (negative) on the complex set of vectors $\mathcal{M} \subset \mathbb{C}^n$ if $x^* S x > 0$ ($x^* S x < 0$) holds for all $x \in \mathcal{M} \setminus \{0\}$. Finally, we always identify a system $\dot{x} = Ax + Bu$, $z = Cx + Du$ with the corresponding Rosenbrock matrix

$$S(s) = \begin{pmatrix} A - sI & B \\ C & D \end{pmatrix}.$$

If $H(s)$ is a real rational matrix, we define $\|H(s)\|_{\infty} := \sup_{\omega \in \mathbb{R}} \|H(i\omega)\| \leq \infty$. $\|H(s)\|_{\infty}$ coincides with the L_{∞} -norm if $H(s)$ is proper and has no poles on the imaginary axis. If $H(s)$ is proper and stable, $\|H(s)\|_{\infty}$ is the H_{∞} -norm of $H(s)$.

2. H_{∞} -optimization and Riccati inequalities. The system is described by

$$(1) \quad \begin{aligned} \dot{x} &= Ax + Bu + Gd, & x(0) &= 0, \\ z &= Hx + Eu, \end{aligned}$$

where $x \in \mathbb{R}^n$ is the state, $u \in \mathbb{R}^m$ the control, $d \in \mathbb{R}^r$ the external disturbance, and $z \in \mathbb{R}^k$ the controlled output, and the real matrices A, B, G, H, E are of compatible size.

The H_{∞} -problem by static state-feedback consists of minimizing $\|(H + EF) \times (sI - A - BF)^{-1}G\|_{\infty}$ over all feedback matrices F such that $A + BF$ is stable. Obviously, $(A - sI \ B)$ must be stabilizable. In this paper, we treat only the *regular* problem and hence we assume in addition that E has maximal column. We can then perform the preliminary feedback $u = Fx + v$ with $F := -(E^T E)^{-1} E^T H$ to achieve $E^T (H + EF) = 0$. The subsequent input coordinate change $v = (E^T E)^{-1/2} u$ shows that we can start (just to simplify the formulae) without restriction with the following requirements:

$$(2) \quad (A - sI \ B) \text{ is stabilizable and } E^T (H \ E) = (0 \ I).$$

Apart from § 3, these are the standing assumptions throughout the paper. For any F of correct size, we introduce the notation

$$\mu(F) := \frac{1}{\|(H + EF)(sI - A - BF)G\|_{\infty}^2}$$

under the usual conventions $1/0 = \infty$, $1/\infty = 0$. The feedback matrix F is called *stabilizing* if $A + BF$ is stable. We study in this paper the H_∞ -optimization problem

$$\mu_* = \sup \{ \mu(F) \mid F \text{ is stabilizing} \}.$$

By stabilizability of $(A - sI \ B)$, μ_* is positive but could be infinite. The first step consists of giving algebraic characterizations for some real parameter μ to be *suboptimal* in the sense of $\mu < \mu_*$. If μ is actually suboptimal, we discuss the construction of stabilizing feedbacks F with $\mu < \mu(F)$. Based on the suboptimality criteria, we finally give a fast procedure to compute μ_* . The following suboptimality characterization is easily extracted from [10], [11], [28], and [18] and is contained in [19] as a special case of a fundamental result for the possibly singular problem.

THEOREM 1. *Fix any $\mu > 0$. Then μ is suboptimal if and only if there exists some $P > 0$ with*

$$(3) \quad A^T P + PA + P(\mu GG^T - BB^T)P + H^T H < 0.$$

If $P > 0$ satisfies (3), then $F := -B^T P$ is μ -suboptimal in the sense of $\mu < \mu(F)$.

The parameter μ enters this ARI linearly. This is the simple reason why we do not infimize $\|(H + EF)(sI - A - BF)^{-1}G\|_\infty$ but, equivalently, maximize $\mu(F)$.

If we reformulate that $\mu > 0$ is suboptimal if and only if there exists an $X > 0$ with

$$(4) \quad (-A^T)^T X + X(-A^T) - X(H^T)(H^T)^T X - \mu GG^T + BB^T > 0,$$

we can explain the difficulties caused by the \mathbb{C}^0 -zeros [1] of the system

$$(5) \quad \begin{pmatrix} A - sI & B \\ H & E \end{pmatrix}$$

which coincide, by (2), with the unobservable modes of

$$S(s) := \begin{pmatrix} A - sI \\ H \end{pmatrix}.$$

Fix some $\mu > 0$ and suppose first that $S(s)$ has only unobservable modes in \mathbb{C}^+ , i.e., the system $(-A^T - sI \ H^T)$ is stabilizable. We anticipate Theorem 2 to infer that (4) has a positive definite solution if and only if the Riccati equation

$$(6) \quad (-A^T)^T X + X(-A^T) - X(H^T)(H^T)^T X - \mu GG^T + BB^T = 0$$

is solvable and its stabilizing (maximal) solution X_+ ($\sigma(-A^T - H^T H X_+) \subset \mathbb{C}^-$) is positive definite. These conditions can be checked in an algebraic way: See if the Hamiltonian

$$H(\mu) := \begin{pmatrix} -A^T & -H^T H \\ \mu GG^T - BB^T & A \end{pmatrix}$$

has no eigenvalues in \mathbb{C}^0 . If this is true, compute a real matrix $(T_1^T \ T_2^T)^T$ whose columns form a basis of the stable subspace of $H(\mu)$. Then T_1 is invertible and we must see if the symmetric stabilizing solution $X_+ = T_2 T_1^{-1}$ of the ARE (6) is positive definite (see, e.g., [6, § 7.2]).

General direct solvability criteria for the ARI (4) are only available under the assumption that $S(s)$ has no unobservable modes in \mathbb{C}^0 : There exists a real symmetric solution of the ARI (4) if and only if $H(\mu)$ has no eigenvalues in \mathbb{C}^0 [5]. But even under this restrictive assumption, no characterization of the existence of positive definite solutions is available.

Our whole interest centers around the situation $\sigma(S(s)) \cap \mathbb{C}^0 \neq \emptyset$, where the results of [5] are not applicable. Since it is not possible in this case to infer from the solvability of the ARI (4) the solvability of the ARE (6), we cannot work with Riccati equations instead of Riccati inequalities. The usual technique to overcome these difficulties is to look at the perturbed ARE

$$(7) \quad (-A^T)^T X + X(-A^T) - X((H^T)(H^T)^T + \varepsilon^2 I)X - \mu GG^T + BB^T = 0.$$

Since $(-A^T - sI \ H^T \ \varepsilon I)$ is controllable, it is clear that μ is suboptimal if and only if there exists an $\varepsilon > 0$ such that (7) has a positive definite solution. The main difficulty is the suitable choice of $\varepsilon > 0$. Although this perturbation approach has obvious disadvantages, it provides the starting point for our derivation of direct solvability tests for general strict algebraic Riccati inequalities.

3. Solvability criteria for a strict algebraic Riccati inequality. We derive in this section checkable conditions for the solvability of the Riccati inequality

$$A^T X + XA - XBB^T X + Q > 0$$

with some real symmetric or even positive definite X . Here the data matrices $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, and $Q \in \mathbb{S}^n$ are completely arbitrary; in particular we do not assume that $(A - sI \ B)$ is controllable or stabilizable. To avoid confusion we stress that all objects introduced in this section are only used here and are independent of those in the rest of the paper.

We decided to investigate this type of ARI instead of the one appearing in H_∞ -theory since our results are somewhat more natural and easier to formulate with respect to the data A, B, Q and since it is more convenient to compare them with those in the literature. In particular we refer the reader to [5], where the most general solvability test for the strict Riccati inequalities appears: If $(A - sI \ B)$ has no uncontrollable modes on the imaginary axis, the ARI (8) has a solution $X \in \mathbb{S}^n$ if and only if the Hamiltonian (9) has no eigenvalues in \mathbb{C}^0 . In this generality the existence of positive definite solutions has not been characterized up to now.

For notational simplicity let us introduce the Riccati map $R: \mathbb{S}^n \rightarrow \mathbb{S}^n$ by

$$X \rightarrow R(X) := A^T X + XA - XBB^T X + Q.$$

If $(A - sI \ B)$ is stabilizable, it is simple to derive the following rather well-known basic results, which will be instrumental in our subsequent considerations.

THEOREM 2. *Suppose that $(A - sI \ B)$ is stabilizable. Then the following three statements are equivalent:*

- (a) $R(X) > 0$ has a solution $X \in \mathbb{S}^n$.
- (b) $R(X) = 0$ has a solution $X_+ \in \mathbb{S}^n$ such that $A - BB^T X_+$ is stable.
- (c) The Hamiltonian matrix

$$(9) \quad \begin{pmatrix} A & -BB^T \\ -Q & -A^T \end{pmatrix}$$

has no eigenvalues on the imaginary axis.

Suppose that one of these conditions holds true. Then the solution X_+ in (b) is unique. This so-called stabilizing solution X_+ has the maximality properties

$$X \in \mathbb{S}^n: R(X) \geq 0 \Rightarrow X \leq X_+ \quad \text{and} \quad R(X) > 0 \Rightarrow X < X_+.$$

Furthermore, there exists a sequence $X_j \in \mathbb{S}^n$ with $R(X_j) > 0$ that converges to X_+ for $j \rightarrow \infty$.

Proof. For a proof of the equivalences we refer to [5]–[7], [16] and stress that there is no need to use (advanced) techniques from symplectic algebra. The statements

about the uniqueness and the maximality of X_+ are combinations of results from [7], [16], and [26]. We note that $R(X) > 0$ implies $X_+ \geq X$. Therefore, $x^T(X_+ - X)x = 0$ yields $X_+x = Xx$, i.e., $x^TR(X_+)x = x^TR(X)x$, which leads to $x = 0$.

We should, however, prove in more detail that X_+ can be approximated by solutions of the strict Riccati inequality. In the case where the Hamiltonian (9) has no eigenvalues in \mathbb{C}^0 , there exists some $\varepsilon_0 > 0$ such that

$$\begin{pmatrix} A & -BB^T \\ -Q + \varepsilon I & -A^T \end{pmatrix}$$

has no eigenvalues in \mathbb{C}^0 for all $\varepsilon \in [0, \varepsilon_0]$. For these ε , there is a unique $X(\varepsilon) \in \mathbb{S}^n$ with $R(X(\varepsilon)) = \varepsilon I > 0$ such that $A - BB^TX(\varepsilon)$ is stable. By maximality, $X(\cdot)$ is nondecreasing for decreasing ε and bounded from above by $X(0)$, i.e., it converges to some X_0 for $\varepsilon \searrow 0$. By $R(X_0) = 0$ and $\sigma(A - BB^TX_0) \subset \mathbb{C}^- \cup \mathbb{C}^0$, we conclude that X_0 coincides with X_+ [26] and we end up with $X(\varepsilon) \rightarrow X_+$ for $\varepsilon \searrow 0$. \square

If $(A - sI \ B)$ is not necessarily stabilizable, the uncontrollable modes of $(A - sI \ B)$ on the imaginary axis pose the main problem [5]. To be able to apply the results in Theorem 2 and to display the zero structure of $(A - sI \ B)$ in \mathbb{C}^0 , it is most helpful to transform $(A - sI \ B)$ with the help of a nonsingular matrix S to

$$(10) \quad A_S := SAS^{-1} = \begin{pmatrix} A_1 & B_1F_2 & B_1F_3 \\ 0 & A_2 & 0 \\ 0 & 0 & A_3 \end{pmatrix}, \quad B_S := SB = \begin{pmatrix} B_1 \\ 0 \\ 0 \end{pmatrix}$$

such that $(A_1 - sI \ B_1)$ is stabilizable, $\sigma(A_2) \in \mathbb{C}^0$, and $\sigma(A_3) \in \mathbb{C}^+ [14]$. Then the eigenstructure of A_2 determines the zero structure of the pencil $(A - sI \ B)$ on the imaginary axis. The standard stabilizable subspaces of $(A - sI \ B)$ with respect to \mathbb{C}^- and $\mathbb{C}^- \cup \mathbb{C}^0$, which are denoted as

$$\mathcal{V}^-(A - sI \ B) \quad \text{and} \quad \mathcal{V}^-(A - sI \ B) + \mathcal{V}^0(A - sI \ B),$$

have (with an obvious partition) a nice explicit description in these special coordinates:

$$\mathcal{V}^-(A_S - sI \ B_S) = \{x_2 = 0, x_3 = 0\} \quad \text{and} \quad \mathcal{V}^-(A_S - sI \ B_S) + \mathcal{V}^0(A_S - sI \ B_S) = \{x_3 = 0\}.$$

The matrix Q is transformed to $Q_S := S^{-T}QS^{-1}$ and Q_S is partitioned as A_S . If defining $X \rightarrow R_S(X) := A_S^TX + XA_S - XB_SB_S^TX + Q_S$, the transformation of Q is motivated by

$$S^TR_S(X)S = R(S^TXS),$$

which shows that

$$\{X \in \mathbb{S}^n \mid R(X) > 0\} = S^T\{X \in \mathbb{S}^n \mid R_S(X) > 0\}S.$$

Therefore it is possible as well to characterize the existence of some symmetric or positive definite X with $R_S(X) > 0$.

If we partition $X \in \mathbb{S}^n$ and $R_S(X)$ as Q , we easily compute the blocks of $R_S(X)$ as

$$R_1(X) := A_1^TX_1 + X_1A_1 - X_1B_1B_1^TX_1 + Q_1,$$

$$R_{12}(X) := (A_1 - B_1B_1^TX_1)^TX_{12} + X_{12}A_2 + X_1B_1F_2 + Q_{12},$$

$$R_{13}(X) := (A_1 - B_1B_1^TX_1)^TX_{13} + X_{13}A_3 + X_1B_1F_3 + Q_{13},$$

$$R_2(X) := A_2^TX_2 + X_2A_2 - (F_2 - B_1^TX_{12})^T(F_2 - B_1^TX_{12}) + F_2^TF_2 + Q_2,$$

$$R_{23}(X) := A_2^TX_{23} + X_{23}A_3 - (F_2 - B_1^TX_{12})^T(F_3 - B_1X_{13}) + F_2^TF_3 + Q_{23},$$

$$R_3(X) := A_3^TX_3 + X_3A_3 - (F_3 - B_1^TX_{13})^T(F_3 - B_1^TX_{13}) + F_3^TF_3 + Q_3.$$

Although these formulas look complicated, they exhibit the dependence of the blocks of $R_S(X)$ on those of X . It will turn out to be of great importance that $R_1(X)$ only depends on X_1 and $R_{12}(X)$ only on X_1 and X_{12} . Furthermore, varying X_3 only changes the block $R_3(X)$ in $R_S(X)$. Now we are ready to formulate the main result of this paper.

THEOREM 3. (a) *There exists an $X \in \mathbb{S}^n$ with $R_S(X) > 0$ if and only if there is some $Y \in \mathbb{S}^n$ such that*

$$(11) \quad \sigma(A_1 - B_1 B_1^T Y_1) \subset \mathbb{C}^-, \quad A_1^T Y_1 + Y_1 A_1 - Y_1 B_1 B_1^T Y_1 + Q_1 = 0,$$

$$(12) \quad (A_1 - B_1 B_1^T Y_1)^T Y_{12} + Y_{12} A_2 + Y_1 B_1 F_2 + Q_{12} = 0,$$

$$(13) \quad x^*[Q_2 + F_2^T F_2 - (F_2 - B_1^T Y_{12})^T (F_2 - B_1^T Y_{12})]x > 0$$

hold for any (possibly complex) eigenvector x of A_2 .

(b) *There exists a real symmetric $X > 0$ with $R_S(X) > 0$ if and only if some Y exists as in (a) which satisfies, in addition, $Y_1 > 0$.*

All conditions given in the theorem are verifiable. First, we must test the existence of Y_1 that satisfies (11). By stabilizability of $(A_1 - sI \ B_1)$ and Theorem 2, we need only check whether the Hamiltonian

$$\begin{pmatrix} A_1 & -B_1 B_1^T \\ -Q_1 & -A_1^T \end{pmatrix}$$

has no eigenvalues in \mathbb{C}^0 . If Y_1 exists, it is *unique* and easily determined by computing a basis of the stable eigenspace of this Hamiltonian. Then the Sylvester equation (12) has a *unique* solution Y_{12} . Since A_2 is real, we propose to test (13) in the following way: For any $i\omega \in \sigma(A_2)$ with $\text{Re}(\omega) \geq 0$, we compute a complex matrix E whose columns form a basis of the complex subspace $\{x \in \mathbb{C}^{n_2} \mid (A_2 - i\omega I)x = 0\}$ ($A_2 \in \mathbb{R}^{n_2 \times n_2}$) and check whether $E^*[Q_2 + F_2^T F_2 - (F_2 - B_1^T Y_{12})^T (F_2 - B_1^T Y_{12})]E$ is positive definite. The additional condition in (b) can be verified by determining the smallest eigenvalue of the uniquely determined matrix Y_1 . Obviously, the uncontrollable modes of $(A - sI \ B)$ in \mathbb{C}^+ do not influence these criteria.

The first step in proving Theorem 3 consists of verifying the result for a Lyapunov inequality

$$A^T X + XA + Q > 0$$

in the case where A has only eigenvalues in \mathbb{C}^0 . This is in itself an interesting and (to our knowledge) new result. We not only characterize the existence of solutions but also show that they can even be chosen to be arbitrarily large. The key idea for the proof is to perturb the Lyapunov inequality and to investigate the resulting parametrized ARE for a *controllable* system, for which a well-established theory exists.

THEOREM 4. *Suppose that the matrix $A \in \mathbb{R}^{n \times n}$ has only eigenvalues on the imaginary axis.*

(a) *The inequality $A^T X + XA + Q > 0$ has a real symmetric solution X if and only if for any eigenvector x of A , the quadratic form $x^* Q x$ is positive.*

(b) *If one of the equivalent conditions of (a) holds there exists for any $X_0 \in \mathbb{S}^n$ a solution $X \in \mathbb{S}^n$ of the Lyapunov inequality with $X > X_0$.*

Proof of (a). If X is some solution of the Lyapunov inequality, we deduce from $(A - i\omega I)^* X + X(A - i\omega I) + Q > 0$ for any $\omega \in \mathbb{R}$ immediately the necessity of our condition if choosing $i\omega \in \sigma(A)$.

To prove that this obvious necessary condition is in fact sufficient, we first note that there exists a $\delta > 0$ such that the implication

$$(14) \quad (A - i\omega I)x = 0, \|x\| = 1 \Rightarrow x^*(Q - \delta I)x > 0$$

holds true. Define for this fixed δ the matrix $Q_\delta := Q - \delta I$ and consider the ARE

$$(15) \quad A^T X + X A - \varepsilon^2 X^2 + Q_\delta = 0.$$

We show that this ARE has a solution $X \in \mathbb{S}^n$ for some $\varepsilon > 0$. This implies $A^T X + X A + Q = \delta I + \varepsilon^2 X^2 > 0$, i.e., X solves the Lyapunov inequality as desired.

Since $(A - sI \ \varepsilon I)$ is controllable for $\varepsilon > 0$, there is a real symmetric solution of (15) if and only if the frequency domain inequality

$$I + (\varepsilon I)(i\omega I - A)^{-*} Q_\delta (i\omega I - A)^{-1} (\varepsilon I) \geq 0$$

holds for all $i\omega \in \mathbb{C}^0 \setminus \sigma(A)$ [4]. If defining $H(i\omega) = (i\omega I - A)^{-*} Q_\delta (i\omega I - A)^{-1}$, the Hermitian matrix $I + H(i\omega)$ converges to the identity for $|\omega| \rightarrow \infty$ and thus there exists an ω_0 such that $I + H(i\omega) > 0$ for all $\omega \in \mathbb{R}$ with $|\omega| > \omega_0$. This implies $I + \varepsilon^2 H(i\omega) > 0$ for all $\varepsilon \in [0, 1]$ and $|\omega| > \omega_0$. With the *bounded* set of frequencies $F := \{\omega \in [-\omega_0, \omega_0] \mid i\omega \notin \sigma(A)\}$, we need only prove that

$$\exists \varepsilon \in (0, 1] \quad \forall \omega \in F: \quad (i\omega I - A)^*(i\omega I - A) + \varepsilon^2 Q_\delta \geq 0.$$

Suppose that this statement is not true. Then we can construct sequences $\omega_j \in F$ and $x_j \in \mathbb{C}^n$ with $\|x_j\| = 1$ such that

$$(16) \quad x_j^* (i\omega_j I - A)^* (i\omega_j I - A) x_j < -\frac{1}{j} x_j^* Q_\delta x_j$$

holds for all $j \in \mathbb{N}$. By boundedness, it is possible to extract a subsequence such that x_{j_l} converges to some $x_\infty \neq 0$ and ω_{j_l} to an $\omega_\infty \in [-\omega_0, \omega_0]$ for $l \rightarrow \infty$. From (16) we deduce $(i\omega_{j_l} I - A) x_{j_l} \rightarrow 0$ for $l \rightarrow \infty$ since $x_{j_l}^* Q_\delta x_{j_l}$ is bounded. This shows $(i\omega_\infty I - A) x_\infty = 0$ and, therefore, x_∞ is an eigenvector of A . The inequality (16), however, yields $x_{j_l}^* Q_\delta x_{j_l} \leq 0$ and thus $x_\infty^* Q_\delta x_\infty \leq 0$, a contradiction to (14).

Proof of (b). Suppose that $Z \in \mathbb{S}^n$ satisfies $A^T Z + Z A + Q > 0$. Then there exists some $\delta_0 > 0$ such that $A^T Z + Z A + Q - \delta_0 I$ is still positive definite. Now consider the parametrized ARE

$$AP + PA^T + \delta_0 P^2 - \varepsilon I = 0.$$

By standard LQ theory this equation has for any $\varepsilon > 0$ a unique positive definite solution $P(\varepsilon)$. Considering the corresponding LQ problem shows immediately that $P(\varepsilon)$ is nonincreasing for decreasing values of ε and thus $P_0 := \lim_{\varepsilon \rightarrow 0} P(\varepsilon)$ exists, is positive semidefinite and satisfies $AP_0 + P_0 A^T + \delta_0 P_0^2 = 0$. Since the zero matrix is another solution of the latter ARE and since A has all its eigenvalues in $\mathbb{C}^0 \cup \mathbb{C}^+$, 0 is the *maximal* solution of this equation [26] and hence we obtain $P_0 \leq 0$, i.e., $P_0 = 0$. Therefore there exists an $\varepsilon_0 > 0$ with $P(\varepsilon_0)^{-1} > X_0 - Z$ and, of course, $P(\varepsilon_0)^{-1}$ satisfies $A^T P(\varepsilon_0)^{-1} + P(\varepsilon_0)^{-1} A + \delta_0 I = \varepsilon_0 P(\varepsilon_0)^{-2} > 0$. This implies that $X := P(\varepsilon_0)^{-1} + Z$ satisfies $X > X_0$ and, by the choice of δ_0 , $A^T X + X A + Q > 0$ as desired. \square

Now we are able to prove Theorem 3. The proof of necessity essentially proceeds along the following ideas. We perturb B to some $B(\varepsilon)$ such that $(A - sI \ B(\varepsilon))$ is stabilizable and infer that the ARE $A^T X + X A - X B(\varepsilon) B(\varepsilon)^T X + Q = 0$ has a stabilizing solution $X(\varepsilon)$ for all small $\varepsilon > 0$. Then it is possible to show that the $(1, 1)$ and $(1, 2)$ block of $X(\varepsilon)$ converge to the corresponding blocks Y_1 and Y_{12} , which must be constructed. The nonstrict version of (13) would follow immediately. To verify the strict version of this statement we must, in addition, perturb the constant matrix Q .

Proof of necessity in Theorem 3. Suppose that $X \in \mathbb{S}^n$ satisfies $R_S(X) > 0$. Then we define the submatrices

$$X_p := \begin{pmatrix} X_1 & X_{12} \\ X_{12}^T & X_2 \end{pmatrix}, \quad A_p := \begin{pmatrix} A_1 & B_1 F_2 \\ 0 & A_2 \end{pmatrix}$$

and the perturbations

$$B_p(\varepsilon) := \begin{pmatrix} B_1 & 0 \\ 0 & \varepsilon I \end{pmatrix}, \quad Q_p(\delta) := \begin{pmatrix} Q_1 & Q_{12} \\ Q_{12}^T & Q_2 - \delta I \end{pmatrix}$$

to infer $A_p^T X_p + X_p A_p - X_p B_p(0) B_p(0)^T X_p + Q_p(0) > 0$. By continuity, there exist $\varepsilon_0 > 0$ and $\delta_0 > 0$ such that

$$A_p^T X_p + X_p A_p - X_p B_p(\varepsilon) B_p(\varepsilon)^T X_p + Q_p(\delta_0) > 0$$

holds for all $\varepsilon \in [0, \varepsilon_0]$. Since $(A_p - sI \ B_p(\varepsilon))$ is stabilizable for $\varepsilon > 0$, we can choose $X(\varepsilon)$ to be the largest matrix satisfying

$$(17) \quad A_p^T X(\varepsilon) + X(\varepsilon) A_p - X(\varepsilon) B_p(\varepsilon) B_p(\varepsilon)^T X(\varepsilon) + Q_p(\delta_0) = 0.$$

We partition $X(\cdot)$ according to A_p and note that

$$(18) \quad A_p - B_p(\varepsilon) B_p(\varepsilon)^T X(\varepsilon) = \begin{pmatrix} A_1 - B_1 B_1^T X_1(\varepsilon) & B_1 F_2 - B_1 B_1^T X_{12}(\varepsilon) \\ -\varepsilon^2 X_{12}(\varepsilon)^T & A_2 - \varepsilon^2 X_2(\varepsilon) \end{pmatrix}$$

is stable. By maximality, $X(\cdot)$ is nondecreasing for decreasing $\varepsilon \in (0, \varepsilon_0]$. To prove further properties of $X(\cdot)$, we write (17) blockwise and get

$$(19) \quad A_1^T X_1(\varepsilon) + X_1(\varepsilon) A_1 - X_1(\varepsilon) B_1 B_1^T X_1(\varepsilon) + Q_1 - (\varepsilon X_{12}(\varepsilon))(\varepsilon X_{12}(\varepsilon))^T = 0,$$

$$(20) \quad (A_1 - B_1 B_1^T X_1(\varepsilon))^T X_{12}(\varepsilon) + X_{12}(\varepsilon) (A_2 - \varepsilon^2 X_2(\varepsilon)) + X_1(\varepsilon) B_1 F_2 + Q_{12} = 0,$$

$$(21) \quad A_2^T X_2(\varepsilon) + X_2(\varepsilon) A_2 + \tilde{Q}_2 - \varepsilon^2 X_2(\varepsilon)^2 - (F_2 - B_1^T X_{12}(\varepsilon))^T (F_2 - B_1^T X_{12}(\varepsilon)) = 0$$

after introducing $\tilde{Q}_2 := F_2^T F_2 + Q_2 - \delta_0 I$.

Multiplying (21) with ε^2 shows that $P(\varepsilon) := \varepsilon^2 X_2(\varepsilon)$ satisfies

$$(22) \quad A_2^T P(\varepsilon) + P(\varepsilon) A_2 + \varepsilon^2 \tilde{Q}_2 - P(\varepsilon)^2 = \varepsilon^2 (F_2 - B_1^T X_{12}(\varepsilon))^T (F_2 - B_1^T X_{12}(\varepsilon))$$

for $\varepsilon \in (0, \varepsilon_0]$. We now prove $P(\varepsilon) \rightarrow 0$ for $\varepsilon \rightarrow 0$.

Choose $i\omega \in \sigma(A_2)$ and take an arbitrary Jordan chain $x_{-1} = 0$, $(A_2 - i\omega I)x_j = x_{j-1}$ for $j = 0, \dots, l$. If we multiply (22) from the left with x_j^* and from the right with x_j , we obtain

$$x_{j-1}^* P(\varepsilon) x_j + x_j^* P(\varepsilon) x_{j-1} + \varepsilon^2 x_j^* \tilde{Q}_2 x_j \geq \|P(\varepsilon) x_j\|^2.$$

For $j = 0$, we deduce $P(\varepsilon) x_0 \rightarrow 0$. Suppose now that $P(\varepsilon) x_{j-1} \rightarrow 0$. Then we conclude from $x_j^* P(\varepsilon) x_{j-1} \rightarrow 0$ and $x_{j-1}^* P(\varepsilon) x_j \rightarrow 0$ with the help of this inequality $P(\varepsilon) x_j \rightarrow 0$. By induction this property thus holds for all $j = 0, \dots, l$. Since A_2 has only eigenvalues in \mathbb{C}^0 and the Jordan chain was arbitrary, we conclude $P(\varepsilon) \rightarrow 0$ for $\varepsilon \rightarrow 0$.

Equation (22) hence implies $\varepsilon^2 (F_2 - B_1^T X_{12}(\varepsilon))^T (F_2 - B_1^T X_{12}(\varepsilon)) \rightarrow 0$, and thus

$$(23) \quad B_1^T [\varepsilon X_{12}(\varepsilon)] \rightarrow 0$$

for $\varepsilon \rightarrow 0$.

Recall that $X_1(\varepsilon)$ is nondecreasing for decreasing ε . By (19), $X_1(\varepsilon)$ satisfies the inequality $A_1^T X_1(\varepsilon) + X_1(\varepsilon) A_1 - X_1(\varepsilon) B_1 B_1^T X_1(\varepsilon) + Q_1 \geq 0$ and is hence bounded by the stabilizing solution of the corresponding Riccati equation. Therefore $X_1(\varepsilon)$ converges to some $X_1(0)$ for $\varepsilon \rightarrow 0$. This implies, again by (19), that $\varepsilon X_{12}(\varepsilon)$ is in fact bounded.

Now choose some F with $\sigma(A_1 - B_1 B_1^T X_1(0) + B_1 F) \subset \mathbb{C}^-$. From (20) we deduce the equation

$$\begin{aligned} & (A_1 - B_1 B_1^T X_1(\varepsilon) + B_1 F)^T [\varepsilon X_{12}(\varepsilon)] + [\varepsilon X_{12}(\varepsilon)] A_2 \\ & = -\varepsilon Q_{12} - [\varepsilon X_1(\varepsilon)] B_1 F_2 + [\varepsilon X_{12}(\varepsilon)] [\varepsilon^2 X_2(\varepsilon)] + F^T B_1^T [\varepsilon X_{12}(\varepsilon)]. \end{aligned}$$

The right-hand side converges to 0 and $X \rightarrow (A_1 - B_1 B_1^T X_1(0) + B_1 F)^T X + X A_2$ is, by the choice of F , a bijective map with bounded inverse. This already implies $\varepsilon X_{12}(\varepsilon) \rightarrow 0$ for $\varepsilon \rightarrow 0$. Thus $X_1(0)$ is in fact a *solution* of the ARE $A_1^T X + X A_1 - X B_1 B_1^T X + Q_1 = 0$. Since the stable matrix (18) is similar to

$$\begin{pmatrix} A_1 - B_1 B_1^T X_1(0) & \varepsilon(B_1 F_2 - B_1 B_1^T X_{12}(\varepsilon)) \\ -\varepsilon X_{12}(\varepsilon)^T & A_2 - \varepsilon^2 X_2(\varepsilon) \end{pmatrix}$$

(using the similarity transformation $\begin{pmatrix} \varepsilon & 0 \\ 0 & 1 \end{pmatrix}$), we obtain first $\sigma(A_1 - B_1 B_1^T X_1(0)) \subset \mathbb{C}^- \cup \mathbb{C}^0$. Since the ARI $A_1^T X + X A_1 - X B_1 B_1^T X + Q_1 > 0$ is solvable, the corresponding Hamiltonian has no eigenvalues in \mathbb{C}^0 and thus $A_1 - B_1 B_1^T X_1(0)$ must in fact be stable.

Now it is possible to infer from (20) the convergence of $X_{12}(\varepsilon)$ to the unique solution $X_{12}(0)$ of the linear equation

$$(A_1 - B_1 B_1^T X_1(0))^T X + X A_2 + X_1(0) B_1 F_2 + Q_{12} = 0.$$

For any eigenvector x of A_2 we deduce from (21)

$$x^* [F_2^T F_2 + Q_2 - (F_2 - B_1^T X_{12}(\varepsilon))^T (F_2 - B_1^T X_{12}(\varepsilon))] x \geq \delta_0 x^* x$$

and thus, after taking the limit,

$$x^* [F_2^T F_2 + Q_2 - (F_2 - B_1^T X_{12}(0))^T (F_2 - B_1^T X_{12}(0))] x > 0.$$

If $X > 0$ solves $R_S(X) > 0$, we infer in addition that $X_1 > 0$ satisfies $A_1^T X_1 + X_1 A_1 - X_1 B_1 B_1^T X_1 + Q_1 > 0$ and hence conclude $X_1 < Y_1$, i.e., $Y_1 > 0$. \square

Sufficiency is shown by constructing X blockwise where we exploit the above-described particular structure of the Riccati map. We select some X_1 with $R_1(X) > 0$ for any real symmetric X that has X_1 as its $(1, 1)$ block. X_1 is chosen near Y_1 such that it is possible to find X_{12} and X_2 with $R_{12}(X) = 0$ and $R_2(X) > 0$. For the special blocks $X_{13} = 0$ and $X_{23} = 0$, we can find a large X_3 to make $R_3(X)$ arbitrarily large. Since only $R_3(X)$ is influenced by X_3 , we can force $R_S(X)$ to be positive definite. If Y_1 is positive definite, X_1 can be chosen to be positive definite and if X_2 and X_3 are large enough, X itself becomes positive definite. For later use, we construct a whole sequence $X(j)$ of solutions.

Proof of sufficiency in Theorem 3. According to Theorem 2, there exists a sequence $X_1(j)$ which converges to Y_1 for $j \rightarrow \infty$ and which satisfies $A_1^T X_1(j) + X_1(j) A_1 - X_1(j) B_1 B_1^T X_1(j) + Q_1 > 0$. Without restriction, this sequence can be chosen such that $A_1 - B_1 B_1^T X_1(j)$ is stable for all $j \in \mathbb{N}$. Hence there exists a sequence $X_{12}(j)$ which satisfies

$$(A_1 - B_1 B_1^T X_1(j))^T X_{12}(j) + X_{12}(j) A_2 + X_1(j) B_1 F_2 + Q_{12} = 0$$

for all $j \in \mathbb{N}$ and which necessarily converges to Y_{12} for $j \rightarrow \infty$. Therefore there exists a $j_0 \in \mathbb{N}$ such that $x^* [Q_2 + F_2^T F_2 - (F_2 - B_1^T X_{12}(j))^T (F_2 - B_1^T X_{12}(j))] x$ is positive for all eigenvectors x of A_2 and all $j \geq j_0$. We can apply Theorem 4 to infer the existence of a sequence $X_2(j)$ with

$$A_2^T X_2(j) + X_2(j) A_2 + Q_2 + F_2^T F_2 - (F_2 - B_1^T X_{12}(j))^T (F_2 - B_1^T X_{12}(j)) > 0$$

and

$$(24) \quad jI < X_2(j)$$

for all $j \geq j_0$. Now define

$$X(j) := \begin{pmatrix} X_1(j) & X_{12}(j) & 0 \\ X_{12}^T(j) & X_2(j) & 0 \\ 0 & 0 & X_3(j) \end{pmatrix}$$

with some still unspecified sequence $X_3(j)$. We obtain

$$R_S(X(j)) = \begin{pmatrix} R_1(X(j)) & 0 & R_{13}(X(j)) \\ 0 & R_2(X(j)) & R_{23}(X(j)) \\ R_{13}(X(j))^T & R_{23}(X(j))^T & A_3^T X_3(j) + X_3(j)A_3 + Q_3 \end{pmatrix}.$$

By construction, $R_1(X(j))$ and $R_2(X(j))$ are positive definite. Since $X_3(j)$ only influences the (3, 3) block and since $-A_3^T$ is stable, $R_3(X(j))$ reaches any symmetric matrix if varying $X_3(j)$ in the symmetric matrices, without changing the other blocks of $R_S(X(j))$. In particular we can find for any $j \geq j_0$ a symmetric $X_3(j)$ such that $R_S(X(j))$ is positive definite. This already proves the sufficiency part of (a).

It is even possible to define a sequence $X_3(j)$ with $R_S(X(j)) > 0$ and $R_3(X(j)) > Q_3 + jI$. The inequality $A_3^T X_3(j) + X_3(j)A_3 > jI$ leads to

$$(25) \quad X_3(j) > j \int_0^\infty e^{-A_3^T t} e^{-A_3 t} dt,$$

which shows in particular that $X_3(j)$ is positive definite. If we recall (24) and the fact that $X_1(j)$, $X_{12}(j)$ converge, $X(j)$ is positive definite for all large j (Lemma 14). \square

Until now the results were given with respect to the transformed data A_S , B_S , and Q_S . We want to propose a possibility to reformulate them in a way which is invariant under the transformation $(A, B, Q) \rightarrow (A_S, B_S, Q_S)$.

First, we characterize the existence of $Y \in \mathbb{S}^n$ such that (11) and (12) hold. This is equivalent to saying that $R_S(Y)$ admits the block triangular shape

$$\begin{pmatrix} 0 & 0 & * \\ 0 & * & * \\ * & * & * \end{pmatrix}$$

and the (1, 1) block $A_1 - B_1 B_1^T Y_1$ of $A_S - B_S B_S^T Y$ is stable. Hence $Y \in \mathbb{S}^n$ satisfies (11) and (12) if and only if $x^T R_S(Y) y$ vanishes for $x \in \mathcal{V}^-(A_S - sI \ B_S)$ and $y \in \mathcal{V}^-(A_S - sI \ B_S) + \mathcal{V}^0(A_S - sI \ B_S)$ and the restriction of $(A_S - B_S B_S^T Y)$ to $\mathcal{V}^-(A_S - sI \ B_S)$ is stable.

DEFINITION 5. \mathcal{T} denotes the set of all matrices $Z \in \mathbb{S}^n$ such that $x^T R(Z) y$ vanishes for all $x \in \mathcal{V}^-(A - sI \ B)$, $y \in \mathcal{V}^-(A - sI \ B) + \mathcal{V}^0(A - sI \ B)$ and such that the restriction of $(A - B B^T Z)$ to $\mathcal{V}^-(A - sI \ B)$ is stable.

If \mathcal{T}_S denotes the corresponding set for (A_S, B_S, Q_S) , we easily derive $\mathcal{T} = S^T \mathcal{T}_S S$. The explicit description

$$\mathcal{T}_S = \left\{ \begin{pmatrix} Y_1 & Y_{12} & * \\ Y_{12}^T & * & * \\ * & * & * \end{pmatrix} \in \mathbb{S}^n \mid Y_1, Y_{12} \text{ are the unique matrices that satisfy (11), (12)} \right\}$$

shows that \mathcal{T} is, if nonempty, a linear manifold in \mathbb{S}^n .

Second, we must reformulate (13). For this purpose we search a complex set $\mathcal{E}(A - sI \ B)$ of \mathbb{C}^n -vectors such that all possible second components of $x \in \mathcal{E}(A_S - sI \ B_S)$ are given by all eigenvectors of A_2 . Obviously, $(A_S - i\omega I)x + B_S u = 0$ implies $(A_2 - i\omega I)x_2 = 0$ and $x_3 = 0$. If x_2 satisfies $(A_2 - i\omega I)x_2 = 0$, we may define

$x := (0 \ x_2^* \ 0)^*$ and $u := -F_2 x_2$ to obtain $(A_S - i\omega I)x + B_S u = 0$. Therefore we define for any $\lambda \in \mathbb{C}$ the *complex* subspace

$$\mathcal{V}^\lambda(A - sI \ B) := \{x \in \mathbb{C}^n \mid \exists u \in \mathbb{C}^m : (A - \lambda I)x + Bu = 0\}$$

and easily verify $\mathcal{V}^\lambda(A_S - sI \ B_S) = S\mathcal{V}^\lambda(A - sI \ B)$. Note that $\mathcal{V}^\lambda(A - sI \ B)$ is the whole \mathbb{C}^n if $(A - sI \ B)$ is controllable [12, Hilfssatz 5.3]. Moreover, the second component of $x \in \mathcal{V}^\lambda(A_S - sI \ B_S)$ is nontrivial (and hence an eigenvector of A_2) if and only if $x \notin \mathcal{V}^-(A_S - sI \ B_S)$.

If we now fix some $Y \in \mathcal{T}_S$, we observe $x^* R_S(Y)x = x_2^* [Q_2 + F_2^T F_2 - (F_2 - B_1^T Y_{12})^T (F_2 - B_1^T Y_{12})] x_2$ for any $x \in \bigcup_{\lambda \in \mathbb{C}^0} \mathcal{V}^\lambda(A_S - sI \ B_S)$ and $x^T Y x = x_1^T Y_1 x_1$ for any $x \in \mathcal{V}^-(A_S - sI \ B_S)$. It is not difficult to verify that these quadratic forms are invariant under the transformation $(A, B, Q) \rightarrow (A_S, B_S, Q_S)$. Therefore we are in the position to reformulate the conditions of Theorem 3 in terms of the original data (A, B, Q) .

THEOREM 6. (a) *There exists an $X \in \mathcal{S}^n$ with $R(X) > 0$ if and only if there exists some $Z \in \mathcal{T}$ such that*

$$x^* R(Z)x > 0 \quad \text{for all } x \in \bigcup_{\lambda \in \mathbb{C}^0} \mathcal{V}^\lambda(A - sI \ B) \setminus \mathcal{V}^-(A - sI \ B).$$

(b) *There exists a $X > 0$ with $R(X) > 0$ if and only if there exists some $Z \in \mathcal{T}$ with $x^* R(Z)x > 0$ for all $x \in \bigcup_{\lambda \in \mathbb{C}^0} \mathcal{V}^\lambda(A - sI \ B) \setminus \mathcal{V}^-(A - sI \ B)$ and*

$$x^* Z x > 0 \quad \text{for all } x \in \mathcal{V}^-(A - sI \ B) \setminus \{0\}.$$

If these conditions hold for *one* $Z \in \mathcal{T}$ then they are valid for all other elements $Z \in \mathcal{T}$.

Let us finally take a closer look at the solution set

$$(26) \quad \mathcal{X} := \{X \in \mathcal{S}^n \mid X > 0, R(X) > 0\}$$

if it is nonempty. If $(A - sI \ B)$ is stabilizable, \mathcal{X} has a strict upper bound $X_+ > 0$, which is, in addition, a limit point of \mathcal{X} (Theorem 2). We can extract from the proof of Theorem 3 that \mathcal{X} has in general no upper bound. Instead we could try to find lower bounds of \mathcal{X}^{-1} . If $(A - sI \ B)$ is stabilizable, $P := X_+^{-1}$ is in fact a *strict lower limit point* of \mathcal{X}^{-1} (Theorem 2) and satisfies $AP + PA^T - BB^T + PQP = 0$ with $\sigma(A + PQ) \subset \mathbb{C}^+$. This important result can be generalized without any assumption on $(A - sI \ B)$.

THEOREM 7. *If \mathcal{X} is not empty, it has the strict lower limit point*

$$(27) \quad P := S^{-1} \begin{pmatrix} Y_1^{-1} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} S^{-T},$$

where $Y_1 > 0$ satisfies (11). P is a positive semidefinite solution of

$$AP + PA^T - BB^T + PQP = 0$$

with $\sigma(A + PQ) \subset \mathbb{C}^0 \cup \mathbb{C}^+$, where the eigenvalues of $A + PQ$ in \mathbb{C}^0 are the uncontrollable modes of $(A - sI \ B)$ in \mathbb{C}^0 (counted with multiplicities).

Proof. It is obviously enough to prove these results for $S = I$. We already saw during the proof of the necessity part of Theorem 3 that $R_S(X) > 0$ implies $X_1 < Y_1$ and Lemma 14 shows $P < X^{-1}$. In the sufficiency part, we constructed $X(j)$, which is positive definite and satisfies $R(X(j)) > 0$ for all large j . Moreover, we infer from (24) $X_2(j)^{-1} \rightarrow 0$ and from (25) $X_3(j)^{-1} \rightarrow 0$ for $j \rightarrow \infty$. Since $X_1(j)^{-1}$ converges to Y_1^{-1} and $X_{12}(j)$ is bounded for $j \rightarrow \infty$, we conclude $X(j)^{-1} \rightarrow P$ for $j \rightarrow \infty$ by Lemma 14.

The other properties of P are obvious. \square

Remark. Suppose that \mathcal{T} is positive on $\mathcal{V}^-(A - sI \ B)$, which is weaker than $\mathcal{X} \neq \emptyset$. Since only the blocks $Y_1 > 0$ and Y_{12} are fixed in the above-given explicit description of \mathcal{T} , this set contains positive definite elements, and P as defined by (27) is a lower limit point of

$$\{Z \in \mathcal{T} \mid Z > 0\}.$$

Moreover, P still has all the properties as listed in Theorem 7. It can be shown that P is in fact the *minimal* under all matrices $X \in \mathbb{S}^n$ which satisfy

$$AX + XA^T - BB^T + XQX = 0, \quad \sigma(A + XQ) \subset \mathbb{C}^0 \cup \mathbb{C}^+, \quad X \geq 0.$$

4. Solution of the H_∞ -optimization problem. We recall from § 2 that $\mu > 0$ is suboptimal if and only if (4) has a positive definite solution. Hence we need only translate the results of § 3 to the present situation.

We first formulate a testable suboptimality criterion in special coordinates that gives the best insight into how the optimal value is restricted. As in § 3, there exists a nonsingular S with

$$S^{-T}AS^T = \begin{pmatrix} A_1 & 0 & 0 \\ K_2H_1 & A_2 & 0 \\ K_3H_1 & 0 & A_3 \end{pmatrix}, \quad S^{-T}G = \begin{pmatrix} G_1 \\ G_2 \\ G_3 \end{pmatrix}, \quad S^{-T}B = \begin{pmatrix} B_1 \\ B_2 \\ B_3 \end{pmatrix},$$

$$HS^T = (H_1 \ 0 \ 0)$$

such that (A_1^{-sI}) has only unobservable modes in \mathbb{C}^+ , $\sigma(A_2) \subset \mathbb{C}^0$, and $\sigma(A_3) \subset \mathbb{C}^-$. Moreover, we denote the set of eigenvalues of A_2 with nonnegative real part as $\{i\omega_1, \dots, i\omega_l\}$ and choose, for any $j \in \{1, \dots, l\}$, complex matrices E_j whose columns form a basis of the eigenspace $\{x \in \mathbb{C}^{n_2} \mid (A_2^T - i\omega_j I)x = 0\}$ (where n_2 is the dimension of A_2).

We immediately arrive at the following central result of our paper which is, in fact, just a corollary to Theorem 3.

THEOREM 8. *Suppose $\mu > 0$. Then μ is suboptimal if and only if there exist matrices $X(\mu)$ and $Y(\mu)$ with*

$$(28) \quad A_1X(\mu) + X(\mu)A_1^T + X(\mu)H_1^TH_1X(\mu) + \mu G_1G_1^T - B_1B_1^T = 0,$$

$$(29) \quad \sigma(A_1 + X(\mu)H_1^TH_1) \subset \mathbb{C}^+,$$

$$(30) \quad X(\mu) > 0,$$

$$(31) \quad (A_1 + X(\mu)H_1^TH_1)Y(\mu) + Y(\mu)A_2^T + X(\mu)H_1^TK_2^T + \mu G_1G_2^T - B_1B_2^T = 0,$$

$$(32) \quad E_j^*[(K_2^T + H_1Y(\mu))^T(K_2^T + H_1Y(\mu)) + \mu G_2G_2^T - B_2B_2^T - K_2K_2^T]E_j < 0$$

for all $j \in \{1, \dots, l\}$.

From the discussion in § 3 it is clear how to check all these conditions algebraically. We stress again that $X(\mu)$ and $Y(\mu)$ are, if existent, uniquely determined, which justifies considering them as functions of μ .

To determine the optimal value μ_* , we must find out when one of these conditions fails to hold. It is a central observation that the solution $X(\mu)$ of the parameter-dependent ARE (28) with property (29) is very well understood [18, §§ V, VI]. In particular there exists a formula for the domain of definition of the function $X(\cdot)$: Compute the unique $X(0)$ which exists and is positive definite by $0 < \mu_*$. Then there exists an $X(\mu)$ with (28) and (29) if and only if

$$\mu < \mu_{\max} := \|H_1(sI - A_1 - X(0)H_1^TH_1)^{-1}G_1\|_\infty^{-2} \leq \infty.$$

Moreover, $X(\cdot)$ is analytic on $(-\infty, \mu_{\max})$ with $X'(\mu) \leq 0$ and $X''(\mu) \leq 0$. Therefore

$$Y(\cdot) \text{ is defined and analytic on } (-\infty, \mu_{\max}).$$

In [18, § VIII] we intensively discussed how to find the second critical parameter

$$\mu_{\text{pos}} := \sup\{\mu < \mu_{\max} \mid X(\mu) > 0\},$$

which is related to the restriction (30). The third critical parameter

$$\mu_{\text{neg}} := \sup\{\mu < \mu_{\max} \mid (32) \text{ holds for all } j = 1, \dots, l\}$$

displays the new restricting condition caused by the \mathbb{C}^0 -zeros of the plant. We immediately deduce that

$$\mu_* = \min\{\mu_{\text{pos}}, \mu_{\text{neg}}\}.$$

In § 5 we give an algorithm for computing μ_{neg} . We will clarify in [19] that μ_{neg} equals $\sup\{\|(H + EF)(sI - A - BF)^{-1}G\|_\infty \mid \sigma(A + BF) \cap \mathbb{C}^0 = \emptyset\}$ if $(A - sI \ B)$ is only assumed to be stabilizable with respect to \mathbb{C}^0 . Therefore μ_{neg} may be interpreted as the optimal value of an L_∞ -optimization problem.

Let us briefly discuss how the various plant zeros influence the optimal value. We can assume without restriction that

$$(33) \quad A_1 = \begin{pmatrix} A_1^o & 0 \\ K_1^o H_1^o & A_1^+ \end{pmatrix}, \quad H_1 = (H_1^o \ 0), \quad G_1 = \begin{pmatrix} G_1^o \\ G_1^+ \end{pmatrix}, \quad B_1 = \begin{pmatrix} B_1^o \\ B_1^+ \end{pmatrix},$$

where

$$\begin{pmatrix} A_1^o - sI \\ H_1^o \end{pmatrix}$$

is observable and $\sigma(A_1^+)$ contains the zeros of $S(s)$ in \mathbb{C}^+ . $X(\mu)$ can be partitioned as

$$\begin{pmatrix} X^o(\mu) & X^{o+}(\mu) \\ X^{o+}(\mu)^T & X^+(\mu) \end{pmatrix}$$

and (28), (29) are equivalent to

$$(34) \quad A_1^o X^o(\mu) + X^o(\mu)(A_1^o)^T + X^o(\mu)(H_1^o)^T H_1^o X^o(\mu) + \mu G_1^o (G_1^o)^T - B_1^o (B_1^o)^T = 0,$$

$$(35) \quad \sigma(A_1^o + X^o(\mu)(H_1^o)^T H_1^o) \subset \mathbb{C}^+$$

plus two *linear* equations for $X^{o+}(\mu)$ and $X^+(\mu)$.

The existence of (the unique) $X^o(\mu)$ with (34) and (35) is equivalent to the existence of $X(\mu)$, i.e., to $\mu < \mu_{\max}$. If $\mu > 0$ is suboptimal, $X^o(\mu)$ exists and is positive definite. This part of the conditions is not related to the zeros of $S(s)$ but is in fact due to $H \neq 0$ and the resulting “observable part” of $S(s)$.

If the plant has \mathbb{C}^+ -zeros, the linear equations which determine $X^{o+}(\mu)$ and $X^+(\mu)$ always have unique solutions for $\mu < \mu_{\max}$. The additional condition caused by these zeros, however, is not only $X^+(\mu) > 0$, but the correct restriction is also coupled to $X^o(\mu)$ via $X^{o+}(\mu)$: $X(\mu) > 0$. This is different for the \mathbb{C}^0 -zeros of $S(s)$. Let us partition $Y(\mu)$ as

$$\begin{pmatrix} Y^o(\mu) \\ Y^+(\mu) \end{pmatrix}.$$

By $H_1 Y(\mu) = H_1^o Y^o(\mu)$, (32) is only influenced by $Y^o(\mu)$, which satisfies

$$[A_1^o + X^o(\mu)(H_1^o)^T H_1^o] Y^o(\mu) + Y^o(\mu) A_2^T + X^o(\mu)(H_1^o)^T K_2^T + \mu G_1^o G_2^T - B_1^o B_2^T = 0.$$

Therefore only $X^o(\mu)$ directly influences the right-hand side of (32), which is in particular not affected by the \mathbb{C}^+ -zero structure! Moreover, (32) is not again coupled to $X^o(\mu)$. The \mathbb{C}^- -zeros do not influence the optimal value at all.

We stress that $G_1^o \neq 0$ causes the genuine difficulties in the H_∞ -problem: If G_1^o vanishes (which is equivalent to $\mu_{\max} = \infty$), $X^o(\cdot)$ and $Y^o(\cdot)$ are *constant*, and $X(\mu)$ as well as the right-hand side of (32) are affine in μ .

What can be said for the feedback construction? Motivated by Theorem 1, we translate Theorem 7 to the present situation and include certain obvious smoothness and monotonicity properties inherited from $X(\cdot)$.

THEOREM 9. (a) For $\mu < \mu_{\text{pos}}$, the matrix

$$P(\mu) = S^{-1} \begin{pmatrix} X(\mu)^{-1} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} S^{-T}$$

satisfies $A^T P(\mu) + P(\mu)A + P(\mu)(\mu GG^T - BB^T)P(\mu) + H^T H = 0$ with $\sigma(A + \mu GG^T P(\mu) - BB^T P(\mu)) \subset \mathbb{C}^- \cup \mathbb{C}^0$, where the eigenvalues in \mathbb{C}^0 are the zeros of $S(s)$ in \mathbb{C}^0 (counted with multiplicities). The kernel of $P(\mu)$ is constant and equal to the undetectable subspace of $S(s)$ with respect to \mathbb{C}^+ . The function $\mu \rightarrow P(\mu)$ is analytic on $(-\infty, \mu_{\text{pos}})$ with $P'(\cdot) \geq 0$ and $P''(\cdot) \geq 0$.

(b) For $\mu < \mu_*$, $P(\mu)$ is the strict lower limit point of

$$(36) \quad \{P > 0 \mid A^T P + PA + P(\mu GG^T - BB^T)P + H^T H < 0\}.$$

We extract the existence of a sequence P_j in (36) with $P_j \rightarrow P(\mu)$ for $j \rightarrow \infty$. Let us define the feedbacks $F_j := -B^T P_j$ and $F_\infty := -B^T P(\mu)$. Then any F_j is stabilizing with $\mu < \mu(F_j)$ (Theorem 1). $A + BF_\infty$ is in general *not* stable and $\sigma(A + BF_\infty) = \sigma(A_1 - B_1 B_1^T X(\mu)^{-1}) \cup \sigma(A_2) \cup \sigma(A_3)$ shows that it has precisely the \mathbb{C}^0 -zeros of $S(s)$ as eigenvalues which are not stabilized. These are, however, canceled in the closed-loop transfer matrix and $H(sI - A - BB^T P(\mu))^{-1} G = H_1(sI - A - B_1 B_1^T X(\mu)^{-1}) G_1$ implies $\mu < \mu(F_\infty)$ [18, Thm. 5]. We summarize these observations by saying that the zeros of $S(s)$ on the imaginary axis generally lead to a jump of the optimal value from μ_{pos} to $\min\{\mu_{\text{pos}}, \mu_{\text{neg}}\}$ but only cause *arbitrarily small* additional feedback.

If $S(s)$ has no \mathbb{C}^0 -zeros, the suboptimality of $\mu > 0$ is characterized by the existence of $X(\mu)$ with (28)–(30). This is equivalent to the existence of $P \geq 0$ with $A^T P + PA + P(\mu GG^T - BB^T)P + H^T H = 0$ and $\sigma(A + \mu GG^T P - BB^T P) \subset \mathbb{C}^-$ since P , if it exists, is unique and coincides with $P(\mu)$. Then $F_\infty = -B^T P$ is stabilizing and still satisfies $\mu < \mu(F)$. We have rederived the results of [3], [24] for \mathbb{C}^0 -zero-free plants.

To design suboptimal feedbacks, we stress that it is neither useful nor necessary to solve the full-order strict ARI (4). Fix $\mu < \mu_*$. We need only compute $X(\mu)$ (by solving a reduced-order Riccati equation), $Y(\mu)$ (by solving a reduced-order linear equation) and to find a solution Z of the reduced-order Lyapunov inequality

$$A_2 Z + Z A_2^T + (K_2^T + H_1 Y(\mu))^T (K_2^T + H_1 Y(\mu)) + \mu G_2 G_2^T - B_2 B_2^T - K_2 K_2^T < 0$$

such that

$$X_p := \begin{pmatrix} X(\mu) & Y(\mu) \\ Y(\mu)^T & Z \end{pmatrix}$$

is positive definite. By [18, Thm. 5], it is obvious that

$$F := -B^T S^{-1} \begin{pmatrix} X_p^{-1} & 0 \\ 0 & 0 \end{pmatrix} S^{-T}$$

is stabilizing and yields the nonstrict inequality $\mu \leq \mu(F)$, which suffices for practical purposes.

Let us finally translate the coordinate invariant formulation appearing in § 3 to the H_∞ -problem. Again this proceeds by simple algebraic dualization and we can be succinct. The main motivation for including this reformulation is the possibility to generalize it literally to the situation when E does not have full column rank [19].

We denote the undetectable subspaces of $S(s)$ with respect to \mathbb{C}^+ or \mathbb{C}^0 by

$$\mathcal{S}_+ \quad \text{or} \quad \mathcal{S}_0,$$

respectively. We further introduce

$$\mathcal{S}_\lambda := \left\{ x \in \mathbb{C}^n \mid \exists u \in \mathbb{C}^n : \begin{pmatrix} x \\ 0 \end{pmatrix} = \begin{pmatrix} A - \lambda I \\ H \end{pmatrix} u \right\},$$

which is motivated by the easily verified duality relation $\mathcal{S}_\lambda = \mathcal{V}^{-\bar{\lambda}}(-A^T - sI \ H^T)^\perp$ for any $\lambda \in \mathbb{C}$. Note that \mathcal{S}_λ is trivial if $S(s)$ is observable.

According to the Riccati map

$$R(X, \mu) := AX + XA^T + XH^THX + \mu GG^T - BB^T$$

defined on $\mathbb{S}^n \times \mathbb{R}$, we introduce for any $\mu \in \mathbb{R}$, the set

$$\mathcal{T}(\mu) := \{ Z \in \mathbb{S}^n \mid ((A + ZH^TH)^T \mid \mathcal{S}_+^\perp) \subset \mathbb{C}^+, x^TR(Z, \mu)y = 0 \text{ for } x \in \mathcal{S}_+^\perp, y \in \mathcal{S}_+^\perp + \mathcal{S}_0^\perp \}$$

and clearly obtain the explicit description

$$\mathcal{T}(\mu) = S^T \left\{ \begin{pmatrix} X(\mu) & Y(\mu) & * \\ Y(\mu)^T & * & * \\ * & * & * \end{pmatrix} \in \mathbb{S}^n \mid X(\mu), Y(\mu) \text{ satisfy (28), (29), (31)} \right\} S.$$

THEOREM 10. *The parameter $\mu > 0$ is suboptimal if and only if there exists a $Z \in \mathcal{T}(\mu)$ such that Z is positive on \mathcal{S}_+^\perp and $R(Z, \mu)$ is negative on $\bigcup_{\lambda \in \mathbb{C}^0} \mathcal{S}_\lambda^\perp \setminus \mathcal{S}_+^\perp$.*

We recall the intuitive interpretation of these conditions: $\mathcal{T}(\mu) \neq \emptyset$ means that a certain lower-dimensional Riccati equation has a antistabilizing solution. We can parametrize $\mathcal{T}(\mu)$ by solving this ARE and a reduced-order linear equation. The positivity condition is just the positive definiteness of the ARE solution. The negativity condition expresses the solvability of a certain reduced-order Lyapunov inequality which is built up with the solution of the linear equation.

Remark. Obviously, $\mathcal{T}(\mu)$ is nonempty if and only if $\mu < \mu_{\max}$. Using the explicit formula for μ_{\max} , it is easily seen that

$$\mu_{\max} = \|C(sI + A + ZH^TH)^{-1}G\|_\infty^{-2}$$

holds for any $Z \in \mathcal{T}(0)$.

5. Computation of the optimal value. We use all the objects defined in § 4 and, in addition, introduce for $j \in \{1, \dots, l\}$ the Hermitian-valued analytic functions

$$F_j(\mu) := -E_j^*[(K_2^T + H_1 Y(\mu))^T (K_2^T + H_1 Y(\mu)) + \mu G_2 G_2^T - B_2 B_2^T - K_2 K_2^T] E_j,$$

$$F(\mu) := \text{blockdiag}(F_1(\mu) \cdots F_l(\mu))$$

on $(-\infty, \mu_{\max})$.

For the computation of μ_* , it remains to determine the critical parameter $\mu_{\text{neg}} = \sup \{ \mu < \mu_{\max} \mid F(\mu) > 0 \}$. To apply the general algorithm in [18, § 7], we must prove $F'(\mu) \leq 0$, $F''(\mu) \leq 0$, and $F(0) > 0$. The last inequality is trivial since $\mu = 0$ is suboptimal. The other two properties are indeed verifiable. Since the proof is nontrivial and

allows us to extract an explicit formula for the derivative $F'(\cdot)$, we present it in detail.

THEOREM 11. *The analytic functions $F_j(\cdot)$ satisfy $F_j(0) > 0$ and $F'_j(\mu) \leq 0$ as well as $F''_j(\mu) \leq 0$ for all $\mu \in (-\infty, \mu_{\max})$ and $j = 1, \dots, l$.*

Collection of formulas. Fix $\mu \in (-\infty, \mu_{\max})$. If defining

$$(H(\mu) \ H_B(\mu) \ H_G(\mu)) := H_1(i\omega_j I + A_1 + X(\mu)H_1^T H_1)^{-1}(X(\mu)H_1^T \ B_1 \ G_1),$$

we can compute $F_j(\mu)$ with the help of

$$(K_2^T + H_1 Y(\mu))E_j = (I - H(\mu))K_2^T E_j - \mu H_G(\mu)G_2^T E_j + H_B(\mu)B_2^T E_j$$

and

$$(37) \ F'_j(\mu) = -E_j^* [G_2 - H_G(\mu)^*(K_2^T + H_1 Y(\mu))]^* [G_2 - H_G(\mu)^*(K_2^T + H_1 Y(\mu))] E_j.$$

Proof. To simplify the exposition we first derive several useful formulas. We fix some $j \in \{1, \dots, l\}$ and define

$$L(\mu) := K_2^T + H_1 Y(\mu) \quad \text{as well as} \quad A(\mu) := (A_1 + X(\mu)H_1^T H_1 + i\omega_j I)^{-1}.$$

The differentiation of (28) leads to

$$(A_1 + X(\mu)H_1^T H_1)X'(\mu) + X'(\mu)(A_1 + X(\mu)H_1^T H_1)^T + G_1 G_1^T = 0,$$

$$(A_1 + X(\mu)H_1^T H_1)X''(\mu) + X''(\mu)(A_1 + X(\mu)H_1^T H_1)^T + 2X'(\mu)H_1^T H_1 X'(\mu) = 0.$$

If we add and subtract $i\omega_j X'(\mu)$ in the first and $i\omega_j X''(\mu)$ in the second equation and multiply the resulting equations with $H_1 A(\mu)$ from the left and with $A(\mu)^* H_1^T$ from the right, we get

$$(38) \quad H_1 X'(\mu)A(\mu)^* H_1^T + H_1 A(\mu)X'(\mu)H_1^T = -[H_1 A(\mu)G_1][H_1 A(\mu)G_1]^*,$$

$$(39) \quad \begin{aligned} & H_1 X''(\mu)A(\mu)^* H_1^T + H_1 A(\mu)X''(\mu)H_1^T \\ & = -2H_1 A(\mu)X'(\mu)H_1^T H_1 X'(\mu)A(\mu)^* H_1^T. \end{aligned}$$

Moreover, we differentiate (31) and derive

$$(A_1 + X(\mu)H_1^T H_1)Y'(\mu) + Y'(\mu)A_2^T + X'(\mu)H_1^T L(\mu) + G_1 G_2^T = 0,$$

$$(A_1 + X(\mu)H_1^T H_1)Y''(\mu) + Y''(\mu)A_2^T + X''(\mu)H_1^T L(\mu) + 2X'(\mu)H_1^T H_1 Y'(\mu) = 0.$$

If we multiply the equations for $Y(\mu)$, $Y'(\mu)$, and $Y''(\mu)$ from the right with E_j and from the left with $H_1 A(\mu)$, we obtain by $A_2^T E_j = i\omega_j E_j$

$$(40) \quad H_1 Y(\mu)E_j = -H_1 A(\mu)[X(\mu)H_1^T K_2^T + \mu G_1 G_2^T - B_1 B_2^T]E_j,$$

$$(41) \quad H_1 Y'(\mu)E_j = -H_1 A(\mu)[X'(\mu)H_1^T L(\mu) + G_1 G_2^T]E_j,$$

$$(42) \quad H_1 Y''(\mu)E_j = -H_1 A(\mu)[X''(\mu)H_1^T L(\mu) + 2X'(\mu)H_1^T H_1 Y'(\mu)]E_j.$$

Equation (40) leads to the formula for $L(\mu)E_j$.

Now we compute by explicit differentiation of $F_j(\cdot)$ and exploiting (41)

$$\begin{aligned} F'_j(\mu) &= -(H_1 Y'(\mu)E_j)^* L(\mu)E_j - E_j^* L(\mu)^T (H_1 Y'(\mu)E_j) - E_j^* G_2 G_2^T E_j \\ &= (L(\mu)E_j)^* [H_1 X'(\mu)A(\mu)^* H_1^T + H_1 A(\mu)X'(\mu)H_1^T] L(\mu)E_j \\ &\quad + E_j^* [G_2 (H_1 A(\mu)G_1)^* L(\mu) + L(\mu)^T (H_1 A(\mu)G_1)G_2^T - G_2 G_2^T] E_j. \end{aligned}$$

After an obvious completion of the squares, we infer

$$\begin{aligned} F'_j(\mu) &= (L(\mu)E_j)^*[H_1X'(\mu)A(\mu)^*H_1^T + H_1A(\mu)X'(\mu)H_1^T]L(\mu)E_j \\ &\quad + (L(\mu)E_j)^*[H_1A(\mu)G_1][H_1A(\mu)G_1]^*L(\mu)E_j \\ &\quad - E_j^*[G_2 - L(\mu)^TH_1A(\mu)G_1][G_2 - L(\mu)^TH_1A(\mu)G_1]^*E_j. \end{aligned}$$

Equation (38) leads to the considerable simplification

$$F'_j(\mu) = -E_j^*[G_2 - L(\mu)^TH_1A(\mu)G_1][G_2 - L(\mu)^TH_1A(\mu)G_1]^*E_j,$$

which is the formula for $F'_j(\mu)$ that we must prove and which shows $F'_j(\mu) \leq 0$.

The second derivative of $F_j(\cdot)$ is given by

$$\begin{aligned} F''_j(\mu) &= -(H_1Y''(\mu)E_j)^*L(\mu)E_j - E_j^*L(\mu)^T(H_1Y''(\mu)E_j) \\ &\quad - 2(H_1Y'(\mu)E_j)^*(H_1Y'(\mu)E_j). \end{aligned}$$

We infer from (42)

$$\begin{aligned} F''_j(\mu) &= E_j^*L(\mu)^T[H_1X''(\mu)A(\mu)^*H_1^T + H_1A(\mu)X''(\mu)H_1^T]L(\mu)E_j \\ &\quad + 2E_j^*[Y'(\mu)^TH_1^TH_1X'(\mu)A(\mu)^*H_1^T]L(\mu) \\ &\quad + L(\mu)^TH_1A(\mu)X'(\mu)H_1^TH_1Y'(\mu)]E_j \\ &\quad - 2E_j^*[Y'(\mu)^TH_1^TH_1Y'(\mu)]E_j^*. \end{aligned}$$

Again a completion of the squares and equation (39) lead to

$$\begin{aligned} F''_j(\mu) &= E_j^*L(\mu)^T[-2H_1A(\mu)X'(\mu)H_1^TH_1X'(\mu)A(\mu)^*H_1^T]L(\mu)E_j \\ &\quad - 2E_j^*[Y'(\mu) - X'(\mu)A(\mu)^*H_1^TL(\mu)]^*H_1^TH_1[Y'(\mu) - X'(\mu)A(\mu)^*H_1^TL(\mu)]E_j \\ &\quad + 2E_j^*[X'(\mu)A(\mu)^*H_1^TL(\mu)]^*H_1^TH_1[X'(\mu)A(\mu)^*H_1^TL(\mu)]E_j \\ &= -2E_j^*[Y'(\mu) - X'(\mu)A(\mu)^*H_1^TL(\mu)]^*H_1^TH_1[Y'(\mu) - X'(\mu)A(\mu)^*H_1^TL(\mu)]E_j, \end{aligned}$$

which implies $F''_j(\mu) \leq 0$. \square

Note that we should avoid computing first the whole matrix $Y(\mu)$ and then $F_j(\mu)$. It is further interesting that $H_1(i\omega_j I + A_1 + X(\mu)H_1^TH_1)^{-1}G_1$ is just the value at $i\omega_j$ of that transfer matrix which plays a central role in the computation of μ_{\max} .

Now we can apply all the results of [18, § 7] since they can be proved in completely the same way for Hermitian-valued functions. We summarize the consequences in the following result.

THEOREM 12. *The critical parameter μ_{neg} equals μ_{\max} if and only if $F(\mu)$ is positive semidefinite for all $\mu \in (-\infty, \mu_{\max})$. Otherwise there exists a $\mu_1 \in (0, \mu_{\max})$ with $F(\mu_1) \not\geq 0$. Then μ_{neg} equals the unique value for which $F(\mu_{\text{neg}})$ is positive semidefinite and singular.*

For a given $\mu_{\text{neg}} \leq \mu_j < \mu_{\max}$ there exists a unique $\mu_{j+1} \in [\mu_{\text{neg}}, \mu_j]$ such that $F(\mu_j) + F'(\mu_j)(\mu_{j+1} - \mu_j)$ is positive semidefinite and singular. The inductively defined sequence μ_j converges monotonically from above and quadratically to μ_{neg} .

Literally the same result holds for the function

$$\mu \rightarrow \text{blockdiag}(X(\mu), F_1(\mu), \dots, F_l(\mu))$$

on $(0, \mu_{\max})$, which allows us to directly compute μ_* instead of determining μ_{neg} and μ_{pos} separately.

Finally we show how to characterize that the critical parameters are infinite, interestingly enough, just by simple algebraic inclusions. Moreover, we stress the important consequences of $\mu_{\max} = \infty$ for the computation of μ_* . We recall from [18] that μ_{\max} is infinite if and only if $X(\cdot)$ is affine and then the parameter μ_{pos} can be determined by solving a symmetric eigenvalue problem. In the same sense the computation of μ_{neg} reduces to the solution of a Hermitian eigenvalue problem.

THEOREM 13. (a) $\mu_{\max} = \infty$ if and only if $\text{im}(G)$ is contained in the unobservable subspace of $S(s)$.

(b) $\mu_{\text{pos}} = \infty$ if and only if $\text{im}(G) \subset \mathcal{S}_+$.

(c) $\mu_{\text{neg}} = \infty$ if and only if $\text{im}(G) \subset \bigcap_{\lambda \in \mathbb{C}^0} \mathcal{S}_\lambda$.

In the case of $\mu_{\max} = \infty$ and $\mu_{\text{neg}} < \infty$, μ_{neg} is the unique value μ for which

$$F(0) - \mu \text{ blockdiag}(E_1^*(G_2 G_2^T)E_1 \cdots E_l^*(G_2 G_2^T)E_l)$$

is positive semidefinite and singular.

Proof. (a) and (b) already appear in [18]. In the case of $\mu_{\max} = \infty$, $H_G(\mu)$ vanishes, and (37) implies $F_j(\mu) = -E_j^* G_2 G_2^T E_j$. Therefore, $F(\cdot)$ is affine and μ_{neg} is finite, with the characterization we have to prove, if and only if there is at least one $j \in \{1, \dots, l\}$ such that $E_j^* G_2 G_2^T E_j$ does not vanish.

We prove (c) by replacing the inclusion with the equivalent fact that $G^T \mathcal{S}_\lambda^\perp$ is trivial for all $\lambda \in \mathbb{C}^0$. Moreover, we can assume (without restriction) $S = I$ and (33). With an obvious notation we infer for $\lambda \in \mathbb{C}^0$ by duality:

$$(43) \quad \mathcal{S}_\lambda^\perp = \{((x_1^0)^T (x_1^+)^T x_2^T x_3^T)^T \mid x_1^+ = 0, (-A_2^T - \lambda I)x_2, x_3 = 0\}.$$

Now $\mu_{\text{neg}} = \infty$ implies $\mu_{\max} = \infty$ and hence $G_1^0 = 0$. Furthermore, $G_2^T E_j = 0$ for any $j \in \{1, \dots, l\}$ shows $(A_2^T - i\omega_j I)x_2 = 0 \Rightarrow G_2^T x_2 = 0$. This obviously leads to $G^T \mathcal{S}_\lambda^\perp = \{0\}$. On the other hand, $G^T \mathcal{S}_\lambda^\perp = \{0\}$ for one $\lambda \in \mathbb{C}^0$ already implies $G_1^0 = 0$, i.e., $\mu_{\max} = \infty$. Therefore, $G^T \mathcal{S}_{i\omega_j}^\perp = \{0\}$ leads to $G_2^T E_j = 0$ for all $j \in \{1, \dots, l\}$. \square

In case of $\mu_{\max} = \infty$, this result displays in a very nice way both qualitatively and quantitatively which parts of and how G influences μ_{neg} and thus restricts the optimal value μ_* . For the still simpler situation $H = 0$, we obtain $F(0) = \text{blockdiag}(E_1^* B_2 B_2^T E_1 \cdots E_l^* B_2 B_2^T E_l)$ and there is no need to solve any ARE. A referee drew our attention to the unpublished paper [8], which contains suboptimality tests for the one block H_∞ -problem by output measurement without restrictions on $\mathbb{C}^0 \cup \{\infty\}$ -zeros. The results of [8] are applicable to our problem if and only if H vanishes and then they boil down to $X(\mu) > 0$ and $F(\mu) > 0$, although both the formulation as well as the derivation (in the frequency domain) are completely different from ours. Even for $H \neq 0$ and $\mu_{\max} = \infty$, Theorem 13 extends the results of [8] but the real difficulties, of course, arise for $\mu_{\max} < \infty$.

Appendix. The following result is a simple consequence of the well-known formula for the inverse of a block matrix.

LEMMA 14. (a) If

$$X = \begin{pmatrix} X_1 & X_{12} \\ X_{12}^T & X_2 \end{pmatrix} \in \mathbb{S}^n$$

is positive definite and $X_1^{-1} > P_1$, the following inequalities hold:

$$X^{-1} \geq \begin{pmatrix} X_1^{-1} & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad X^{-1} > \begin{pmatrix} P_1 & 0 \\ 0 & 0 \end{pmatrix}.$$

(b) Suppose that $X_1(j)$ converges to $X_1 > 0$, $X_2(j)^{-1}$ converges to 0, and $X_{12}(j)$ is bounded for $j \rightarrow \infty$. Then

$$X(j) := \begin{pmatrix} X_1(j) & X_{12}(j) \\ X_{12}(j)^T & X_2(j) \end{pmatrix}$$

is positive definite for all large j and

$$\lim_{j \rightarrow \infty} X(j)^{-1} = \begin{pmatrix} X_1^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

Acknowledgments. I gratefully acknowledge the support of Siep Weiland from the University of Groningen who wasted many hours for very helpful discussions about the treatment of system zeros on the imaginary axis in the almost disturbance decoupling problem. My thanks also go to the other members of the Systems and Control Group in Groningen (and especially to Professor J. C. Willems) for their kind hospitality.

REFERENCES

- [1] H. ALING AND J. M. SCHUMACHER, *A nine-fold canonical decomposition for linear systems*, Internat. J. Control, 39 (1984), pp. 779–805.
- [2] B. D. O. ANDERSON AND S. VONGPANITLERD, *Network Analysis and Synthesis*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
- [3] J. DOYLE, K. GLOVER, P. KHARGONEKAR, AND B. FRANCIS, *State-space solutions to standard H_∞ and H_2 control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831–847.
- [4] L. E. FAIBUSOVICH, *Algebraic Riccati equation and symplectic algebra*, Internat. J. Control, 43 (1986), pp. 781–792.
- [5] ———, *Matrix Riccati inequality: existence of solutions*, Systems Control Lett., 9 (1987), pp. 59–64.
- [6] B. A. FRANCIS, *A Course in H_∞ Control Theory*, Lecture Notes in Control and Information Systems, Springer-Verlag, Berlin, New York, 1987.
- [7] I. 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] S. HARA, T. SUGIE, AND R. KONDO, *Descriptor form solution for H_∞ control problem with $j\omega$ -axis zeros*, submitted.
- [9] E. A. JONCKHEERE AND J. C. JUANG, *Fast computation of achievable feedback performance in mixed sensitivity H^∞ design*, IEEE Trans. Automat. Control, 32 (1987), pp. 896–906.
- [10] P. P. KHARGONEKAR, I. R. PETERSEN, AND M. A. ROTEA, *H_∞ -optimal control with state-feedback*, IEEE Trans. Automat. Control, 33 (1988), pp. 786–788.
- [11] P. P. KHARGONEKAR, I. R. PETERSEN, AND K. ZHOU, *Robust stabilization of uncertain linear systems: quadratic stabilizability and H_∞ control theory*, IEEE Trans. Automat. Control, 35 (1990), pp. 356–361.
- [12] H. W. KNOBLOCH AND H. KWAKERNAAK, *Linear Kontrolltheorie*, Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1985.
- [13] B. P. MOLINARI, *The stabilizing solution of the algebraic Riccati equation*, SIAM J. Control, 11 (1973), pp. 262–271.
- [14] A. S. MORSE, *Structural invariants of linear multivariable systems*, SIAM J. Control, 11 (1973), pp. 446–465.
- [15] I. R. PETERSEN, *Disturbance attenuation and H_∞ optimization: a design method based on the algebraic Riccati equation*, IEEE Trans. Automat. Control, 32 (1987), pp. 427–429.
- [16] A. C. M. RAN AND R. VREUGDENHIL, *Existence and comparison theorems for algebraic Riccati equations for continuous- and discrete-time systems*, Linear Algebra Appl., 99 (1988), pp. 63–83.
- [17] C. SCHERER, *H_∞ -control by state-feedback: An iterative algorithm and characterization of high-gain occurrence*, Systems Control Lett., 12 (1989), pp. 383–391.
- [18] ———, *H_∞ -control by state-feedback and fast algorithms for the computation of optimal H_∞ -norms*, IEEE Trans. Automat. Control, 35 (1990), pp. 1090–1099.
- [19] ———, *H_∞ -optimization without assumptions on finite or infinite zeros*, SIAM J. Control Optim., 30 (1992), this issue, pp. 143–166.

- [20] C. SCHERER, *The solution set of the algebraic Riccati equation and the algebraic Riccati inequality*, Linear Algebra Appl., 153 (1991), pp. 99–122.
- [21] A. A. STOORVOGEL, *H_∞ control with state feedback*, Proc. MTNS-89, Amsterdam, 1989, to appear.
- [22] —, *The H_∞ control problem: A state space approach*, Ph.D. thesis, Eindhoven University of Technology, the Netherlands, 1990.
- [23] A. A. STOORVOGEL AND H. H. TRENTelman, *The singular H_∞ control problem with state-feedback*, SIAM J. Control Optim., 28 (1990), pp. 1190–1208.
- [24] G. TADMOR, *H_∞ in the time domain: the standard four blocks problem*, Math. Controls Signals Systems, to appear.
- [25] J. C. WILLEMS, *Least-squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, 21 (1971), pp. 319–338.
- [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, *Robust stabilization of linear systems with norm bounded time varying uncertainty*, Systems Control Lett., 10 (1988), pp. 17–20.
- [28] —, *An algebraic Riccati equation approach to H_∞ optimization*, Systems Control Lett., 11 (1988), pp. 85–91.

H_∞ -OPTIMIZATION WITHOUT ASSUMPTIONS ON FINITE OR INFINITE ZEROS*

CARSTEN SCHERER†

Abstract. Explicit algebraic conditions are presented for the suboptimality of some parameter in the H_∞ -optimization problem by output measurement control. Apart from two strict properness conditions, no artificial assumptions restrict the underlying system. In particular, the plant may have zeros on the imaginary axis or at infinity. These suboptimality characterizations are applied to show how to compute the optimal value by quadratically convergent algorithms and to solve the almost disturbance decoupling problem with internal stability.

Key words. H_∞ -optimization, invariant zeros, Riccati inequalities, almost disturbance decoupling

AMS(MOS) subject classifications. 93B27, 93C05, 93C35, 93C45, 93C60, 93D15, 49B99

1. Introduction. After Zames introduced the basic motivations for H_∞ -optimization in [41], this problem has attracted much research, and several techniques for its solutions have been proposed. We only want to mention briefly the approaches in the frequency domain using operator theory [8], [9], J -spectral factorization [2], [11], polynomial methods [18], or an all-pass imbedding (including a treatment of the H_∞ -problem at optimality) [20]. The direct solutions of the H_∞ -problem in the state space are primarily based on the bounded real lemma and on ideas from LQ-theory as well as differential games [3]–[5], [14]–[17], [24], [33], [35]. We refer the reader to [8], [9], [5] for rather comprehensive discussions of the historical development.

The paper [5] may be viewed as the breakthrough for the state-space techniques to solve the H_∞ -problem. The suboptimality tests are formulated in terms of the solvability of two indefinite Riccati equations and a coupling condition on their solutions. Some authors tried to remove several artificial assumptions needed in the approach of [5]. In [4] we can find generalized formulae if certain direct feedthrough matrices do not vanish. From a practical point of view, the main restriction results from the hypothesis that the plant cannot have zeros either on the imaginary axis or at infinity. Using tools of the geometric theory, the zero assumption at infinity was removed in [33], and the suboptimality criteria were given in terms of the solvability of two quadratic matrix inequalities and a coupling condition on their solutions. The solutions of these inequalities may be constructed by transforming the system and again solving two reduced-order indefinite Riccati equations. The general problem without any assumption on the zeros on the imaginary axis and at infinity has not been solved up until now.

In [31] we explained the difficulties resulting from the imaginary axis zeros in the regular H_∞ -problem by state feedback: we cannot infer from the solvability of an algebraic Riccati *inequality* (ARI) the solvability of the corresponding algebraic Riccati *equation* (ARE). The same problems arise in the general H_∞ -problem by output feedback. Indeed from the very beginning of the direct state-space approach to the H_∞ -problem [24] it became clear that Riccati *inequalities* could be viewed as a central

* Received by the editors December 29, 1989; accepted for publication (in revised form) December 13, 1990.

† Mathematisches Institut, Am Hubland, D-8700 Würzburg, Germany. The author was supported by Deutscher Akademischer Austauschdienst. This work was conducted while the author visited the Mathematics Institute of the University of Groningen, the Netherlands.

tool for studying this problem. In fact new algebraic tests for the solvability of general strict Riccati inequalities lead to a satisfactory solution of the regular state-feedback H_∞ -problem [31].

In this paper these ideas are extended to attack the general H_∞ -optimization problem by output measurement and without assumptions on finite or infinite zeros of the underlying plant. Our approach only refers to a strict version of the bounded real lemma and proceeds, by simple algebraic manipulations, completely in the state space.

The paper is organized as follows. In § 3, we will be able to provide a very short but complete proof of a characterization of suboptimality in terms of the solvability of two Riccati inequalities and a coupling condition on their solutions where the ARIs still depend on an unknown feedback and injection matrix (Theorem 1). We shortly discuss the results appearing in the literature. For a reformulation of this characterization into algebraic tests, we need to transform the system data, which is explained in § 4. Then we show in § 5 how to check the solvability of the feedback-dependent ARI by the solvability of a *fixed* well-defined reduced-order ARI (Theorem 6). An abstract reformulation (Theorem 10) displays how to express the suboptimality directly in terms of the original system matrices. These considerations lead to the solution of the general H_∞ -optimization problem by state feedback. Section 6 is devoted to clarifying how the coupling condition can be checked algebraically, culminating in the solution of the general H_∞ -problem by output measurement (Theorem 14). Section 7 contains the application of these results to the computation of the optimal value by quadratically convergent algorithms. In addition, we display certain special cases for which it may be determined by solving Hermitian eigenvalue problems. Finally, we derive in § 8 the geometric solution of the almost disturbance decoupling problem with stability.

We adopt all the notations from [31]. Moreover, we denote by A^+ the Moore-Penrose inverse of $A \in \mathbb{R}^{n \times m}$. The symbol \mathbb{C}^g is used for any nonempty subset of \mathbb{C} which is symmetric with respect to the real axis.

2. Problem formulation. The system is described by

$$(1) \quad \begin{aligned} \dot{x} &= Ax + Bu + Gd, & x(0) &= 0, \\ y &= Cx + Dd, \\ z &= Hx + Ed, \end{aligned}$$

where $x \in \mathbb{R}^n$ is the state, $d \in \mathbb{R}^r$ is the external disturbance, $u \in \mathbb{R}^m$ is the control, $z \in \mathbb{R}^k$ is the controlled output, and $y \in \mathbb{R}^p$ is the measured output available for control. A *compensator* is any dynamic output-feedback system

$$(2) \quad \begin{aligned} \dot{w} &= Kw + Ly, & w(0) &= 0, \\ u &= Mw + Ny. \end{aligned}$$

This system is identified with

$$C_r(s) = \begin{pmatrix} K - sI & L \\ M & N \end{pmatrix},$$

where, for clarity, the dimension $r \in \mathbb{N} \cup \{0\}$ ($K \in \mathbb{R}^{r \times r}$) appears as an index. The closed-loop system is given by

$$\left(\begin{array}{cc|c} A + BNC - sI & BM & G + BND \\ LC & K - sI & LD \\ \hline H + ENC & EM & END \end{array} \right)$$

with state $\begin{pmatrix} x \\ w \end{pmatrix}$ and initial value 0. Moreover, the compensator $C_r(s)$ is called *stabilizing* if it internally stabilizes the resulting closed-loop system in the sense of

$$\sigma \begin{pmatrix} A + BNC & BM \\ LC & K \end{pmatrix} \subset \mathbb{C}^-.$$

It is well known that a stabilizing controller exists and only if

$$(A - sI \quad B) \text{ is stabilizable and } \begin{pmatrix} A - sI \\ C \end{pmatrix} \text{ is detectable.}$$

Apart from § 4, these are the standing assumptions throughout this paper.

For any stabilizing compensator $C_r(s)$, let $\gamma(C_r(s))$ denote the H_∞ -norm of the controlled closed-loop system. The general H_∞ -problem by output measurement consists of minimizing $\gamma(C_r(s))$ over all stabilizing controllers $C_r(s)$. For the same reasons as in [31], we rather consider the equivalent problem: Maximize

$$\mu(C_r(s)) := \frac{1}{\gamma(C_r(s))^2}$$

over all stabilizing $C_r(s)$ under the usual convention $1/0 = \infty$. We define the optimal value by

$$\mu_{\text{opt}} := \sup \{ \mu(C_r(s)) \mid C_r(s) \text{ is stabilizing, } r \in \mathbb{N} \}.$$

Now we can describe the subjects of the present paper:

- Give checkable (algebraic) conditions for some parameter $\mu > 0$ to be suboptimal in the sense of $\mu < \mu_{\text{opt}}$.
- Find a fast algorithm for computing μ_{opt} .
- Try to determine when μ_{opt} can be computed explicitly or characterize whether μ_{opt} is infinite.

All these problems will find rather satisfactory answers. The only restriction on the plant is that there is no direct feedthrough from u to y (which is easy to handle) and from d to z (which is more difficult to deal with). For techniques to overcome these assumptions we refer the reader to [26]. A particular case has its own interest: The state-feedback H_∞ -problem that is defined by the specification

$$C = I \quad \text{and} \quad D = 0;$$

i.e., the whole state is available for control and it is not directly corrupted by the disturbances.

In the frequency domain [8], [9], the plant (including weighting matrices) is described by the real rational proper matrix $G(s)$ as

$$\begin{pmatrix} z(s) \\ y(s) \end{pmatrix} = \begin{pmatrix} G_{11}(s) & G_{12}(s) \\ G_{21}(s) & G_{22}(s) \end{pmatrix} \begin{pmatrix} d(s) \\ u(s) \end{pmatrix}.$$

In our case the system is only restricted to be internally stabilizable and to have a representation with *strictly* proper matrices $G_{11}(s)$ and $G_{22}(s)$. In the usual frequency domain approach, we first parametrize the set of transfer matrices of all internally stabilized closed-loop systems as

$$\{ T_1(s) + T_2(s)Q(s)T_3(s) \mid Q(s) \text{ real rational, proper, stable} \}$$

with $T_2(s) = (H + EF)(sI - A - BF)^{-1}B$ (for some F satisfying $\sigma(A + BF) \subset \mathbb{C}^-$), $T_3(s) = C(sI - A - JC)^{-1}(G + JD)$ (for some J satisfying $\sigma(A + JC) \subset \mathbb{C}^-$) and some real rational proper stable $T_1(s)$. This step changes the H_∞ -problem to a *model-matching problem*.

If $T_2(\lambda)$ has maximal column rank and $T_3(\lambda)$ has maximal row rank for all $\lambda \in \mathbb{C}^0 \cup \{\infty\}$, it is possible to transform the model-matching problem into the so-called *four block Nehari problem* [9]. The aim of this paper is to overcome these rank assumptions, which imply that the transformation to the four block problem is not possible anymore.

3. Suboptimality and strict algebraic Riccati inequalities. We start by characterizing the suboptimality of some positive μ by the solvability of Riccati inequalities which still depend on some unknown feedback and injection matrix.

THEOREM 1. $\mu > 0$ is suboptimal and only if there exist F and J as well as $P > 0$ and $Q > 0$, which satisfy

$$(3) \quad (A + BF)^T P + P(A + BF) + \mu P G G^T P + (H + EF)^T (H + EF) < 0,$$

$$(4) \quad (A + JC)Q + Q(A + JC)^T + \mu Q H^T H Q + (G + JD)(G + JD)^T < 0,$$

$$(5) \quad \rho(PQ) < \frac{1}{\mu}.$$

If these conditions are satisfied, there exists a stabilizing controller $C_n(s)$ of dimension n with $\mu < \mu(C_n(s))$.

This result has a very nice interpretation. Suppose that F is any matrix such that (3) has a positive definite solution. Then this feedback F yields $\sigma(A + BF) \subset \mathbb{C}^-$ and $\|(H + EF)(sI - A - BF)^{-1}G\|_\infty < 1/\mu$. On the other hand, if F satisfies these last two properties, there exists a $P > 0$, which solves (3). Therefore there exist F and $P > 0$ with (3) if and only if μ is suboptimal for the H_∞ -problem by *static state feedback* for the plant

$$S(s) := \begin{pmatrix} A - sI & B \\ H & E \end{pmatrix}$$

and the disturbance input matrix G .

Dually, the existence of J and $Q > 0$ with (4) is equivalent to the suboptimality of μ for the H_∞ -estimation problem (by linear observers) for the plant

$$T(s) := \begin{pmatrix} A - sI & G \\ C & D \end{pmatrix}$$

and the output Hx to be estimated [23]. However, for $\mu < \mu_{\text{opt}}$ it does not suffice that μ is suboptimal for these two problems separately; we must require the additional coupling condition (5). This constraint can be understood if trying to build a compensator: The controller may be constructed as an estimator for the desired suboptimal control function Fx (which is not implementable) using only the measurements y . The conditions for this estimation to be possible result in (5) [5].

From a practical point of view Theorem 1 is not very useful: It is not clear how to test these conditions in algebraic way without resorting to perturbation techniques which would involve an additional parameter. Moreover, it does not provide any insight into how the \mathbb{C}^0 -zeros of the plants $S(s)$ or $T(s)$ influence the optimal value μ_{opt} . Contrary to the considerations in [27], which stop at this point, this is only a preliminary step in our approach.

Theorem 1 is proved in [27] (and was derived independently for the regular problem in the first version of this paper). Nevertheless, we would like to include a complete and most direct proof which is still much shorter than that in [27] since there will be no need to distinguish between proper and nonproper controllers. Moreover,

we prove the sufficiency part by starting with a compensator which is only a slight modification of that used for the regular problem in [4], [5]; the modification can be easily motivated and the required computations are very simple.

Proof of necessity. Suppose that $\mu > 0$ is suboptimal. Then there exists a stabilizing controller $C_r(s)$ with $\mu < \mu(C_r(s))$. Let us define the abbreviations

$$\begin{aligned} \mathcal{A} &:= \begin{pmatrix} A+BNC & BM \\ LC & K \end{pmatrix}, & \mathcal{B} &:= \begin{pmatrix} G+BND \\ LD \end{pmatrix}, \\ \mathcal{C} &:= (H+ENC \quad EM), & \mathcal{D} &:= END. \end{aligned}$$

By $\sigma(\mathcal{A}) \subset \mathbb{C}^-$ and $\|\mathcal{C}(sI - \mathcal{A})^{-1}\mathcal{B} + \mathcal{D}\|_\infty^2 < 1/\mu$, we can apply the strict version of the bounded real lemma and infer the existence of some $\mathcal{Y} > 0$ with [27], [42]

$$(6) \quad \begin{pmatrix} \mathcal{A}^T \mathcal{Y} + \mathcal{Y} \mathcal{A} + \mathcal{C}^T \mathcal{C} & \mathcal{Y} \mathcal{B} + \mathcal{C}^T \mathcal{D} \\ \mathcal{B}^T \mathcal{Y} + \mathcal{D}^T \mathcal{C} & \mathcal{D}^T \mathcal{D} - (1/\mu)I \end{pmatrix} < 0.$$

It is very simple to see [27] that

$$(7) \quad \mathcal{X} := \frac{1}{\mu} \mathcal{Y}$$

satisfies

$$(8) \quad \begin{pmatrix} \mathcal{A} \mathcal{X} + \mathcal{X} \mathcal{A}^T + \mathcal{B} \mathcal{B}^T & \mathcal{X} \mathcal{C}^T + \mathcal{B} \mathcal{D}^T \\ \mathcal{C} \mathcal{X} + \mathcal{D} \mathcal{B}^T & \mathcal{D} \mathcal{D}^T - (1/\mu)I \end{pmatrix} < 0.$$

Let us now partition

$$\mathcal{X} = \begin{pmatrix} X & X_{12} \\ X_{12}^T & X_2 \end{pmatrix} \quad \text{and} \quad \mathcal{Y} = \begin{pmatrix} Y & Y_{12} \\ Y_{12}^T & Y_2 \end{pmatrix}$$

according to \mathcal{A} . Note that both matrices in (6) and (8) are in fact partitioned into three block rows/columns.

Now we just compute the (1, 1) block of the matrix in (6) to

$$\begin{aligned} & A^T Y + Y A + C^T [N^T B^T Y + L^T Y_{12}^T + N^T E^T H] + [Y B N + Y_{12} L + H^T E N] C \\ & + H^T H + (ENC)^T ENC, \end{aligned}$$

which equals

$$(A+JC)^T Y + Y(A+JC) + H^T H + (ENC)^T (ENC)$$

for $J := BN + Y^{-1} Y_{12} L + Y^{-1} H^T E N$. The (1, 3) of (6) block may be written with the help of J as

$$Y(G+JD) + (ENC)^T (END).$$

By

$$\begin{pmatrix} (ENC)^T (ENC) & (ENC)^T (END) \\ (END)^T (ENC) & (END)^T (END) \end{pmatrix} \geq 0,$$

we immediately infer from (6) the inequality

$$\begin{pmatrix} (A+JC)^T Y + Y(A+JC) + H^T H & Y(G+JD) \\ (G+JD)^T Y & -(1/\mu)I \end{pmatrix} < 0.$$

The Schur complement of the matrix on the left in this inequality with respect to the (2, 2) block is negative definite. If we multiply it with μ we infer that $Q := (\mu Y)^{-1} > 0$ satisfies (4).

Since (8) is dual to (6), there is *no need for any computation* to deduce that $P := (\mu X)^{-1}$ is a *positive definite* solution to (3) for $F := NC + MX_{12}^T X^{-1} + NDG^T X^{-1}$.

By the well-known formula for the inverse of a block matrix, the left upper block of \mathcal{Y}^{-1} is given by $(Y - Y_{12} Y_2^{-1} Y_{12}^T)^{-1}$. Therefore (7) shows $\mu X \geq Y^{-1}$, which clearly leads, by the definition of P and Q , to $\rho(PQ) \leq 1/\mu$. If the inequality is not strict, we can perturb P or Q (e.g., to $P - \varepsilon I$ or $Q - \varepsilon I$, $\varepsilon > 0$) such that these perturbations still satisfy the ARIs (3) or (4) and, in addition, the strict coupling condition (5) becomes true.

Proof of sufficiency and controller construction. Let R_F and R_J denote the left-hand sides of the Riccati inequalities (3) and (4). We assume without restriction

$$(9) \quad Q^{-1} R_J Q^{-1} < \mu R_F.$$

Suppose that this is not valid. Obviously, there exists some matrix $R < 0$ with $Q^{-1} R_J Q^{-1} < \mu R$ and $R_F < \mu R$. Since the ARI $(A + BF)^T X + X(A + BF) + \mu X G G^T X + (H + EF)^T (H + EF) - \mu R < 0$ has the solution $X = P$ and since $A + BF$ is stable, there exists some $\bar{P} < P$ [31] with $(A + BF)^T \bar{P} + \bar{P}(A + BF) + \mu \bar{P} G G^T \bar{P} + (H + EF)^T (H + EF) - \mu R = 0$. The stability of $A + BF$ and $\mu R < 0$ implies $\bar{P} > 0$. We can hence replace P by \bar{P} such that all the hypotheses in the theorem persist to hold and (9) becomes true.

Motivated by the results for the regular problem [4], [5], we introduce the *strictly proper* compensator

$$\begin{aligned} \dot{w} &= (A + \mu G G^T P + \Delta)w + Bu + \tilde{J}((C + \mu D G^T P)w - y), \quad w(0) = 0, \\ u &= Fw \end{aligned}$$

in observer form, where we define

$$\begin{aligned} \tilde{J} &:= (I - \mu Q P)^{-1} J, \\ \Delta &:= -\mu(Q^{-1} - \mu P)^{-1}[(BF)^T P + (EF)^T (H + EF)]. \end{aligned}$$

The only modification consists of the introduction of Δ . The formula for Δ will immediately become clear by the computations shown below.

As usual, we consider the dynamics of the error $e := x - w$, which is clearly given by

$$\dot{e} = (A + \mu G G^T P + \tilde{J}(C + \mu D G^T P) + \Delta)e - ((G + \tilde{J}D)G^T(\mu P) + \Delta)x + (G + \tilde{J}D)d.$$

This leads to the closed-loop system (with state $(x^T e^T)^T$)

$$\left(\begin{array}{cc|c} A + BF - sI & -BF & G \\ -(G + \tilde{J}D)G^T(\mu P) - \Delta & A + \mu G G^T P + \tilde{J}(C + \mu D G^T P) + \Delta - sI & G + \tilde{J}D \\ \hline H + EF & -EF & 0 \end{array} \right),$$

which is again denoted as

$$\begin{pmatrix} \mathcal{A} - sI & \mathcal{B} \\ \mathcal{C} & 0 \end{pmatrix}.$$

We must show that $\sigma(\mathcal{A}) \subset \mathbb{C}^-$ and $\|\sqrt{\mu}\mathcal{C}(sI - \mathcal{A})^{-1}\mathcal{B}\|_\infty < 1$.

To finish the proof, it suffices (by the bounded real lemma [42]) to construct a solution \mathcal{Y} of the strict ARI

$$(10) \quad \mathcal{A}^T \mathcal{Y} + \mathcal{Y} \mathcal{A} + \mathcal{Y} \mathcal{B} \mathcal{B}^T \mathcal{Y} + \mu \mathcal{C}^T \mathcal{C} < 0.$$

Again motivated by the solution of the regular problem [5], we are led to try

$$\mathcal{Y} := \begin{pmatrix} \mu P & 0 \\ 0 & Q^{-1} - \mu P \end{pmatrix},$$

which is positive definite by (5). Let us introduce the abbreviation $Z := Q^{-1} - \mu P$.

Now we just compute the blocks of the left-hand side of (10). The (1, 1) block is immediately seen to be μR_F , which is negative definite. The (2, 1) block is given by

$$\begin{aligned} & -(BF)^T(\mu P) - Z(G + \tilde{J}D)G^T(\mu P) - Z\Delta + Z(G + \tilde{J}D)G^T(\mu P) - \mu(EF)^T(H + EF) \\ & = -\mu(BF)^TP - \mu(EF)^T(H + EF) - (Q^{-1} - \mu P)\Delta, \end{aligned}$$

which just vanishes by the very definition of Δ . We first stress $Z\tilde{J} = Q^{-1}J$ and then compute the (2, 2) block to

$$\begin{aligned} & (A + \mu GG^TP)^TZ + Z(A + \mu GG^TP) + ZGG^TZ + \mu(EF)^T(EF) + \Delta^TZ + Z\Delta \\ & + Q^{-1}JDD^TJ^TQ^{-1} + (C^T + \mu PGD^T)J^TQ^{-1} + ZGD^TJ^TQ^{-1} \\ & + Q^{-1}J(C + \mu DG^TP) + Q^{-1}JDG^TZ, \end{aligned}$$

which is rewritten, by $\mu PGG^TZ + \mu ZGG^TP + ZGG^TZ = Q^{-1}GG^TQ^{-1} - (\mu P)GG^T(\mu P)$, as

$$\begin{aligned} & A^TZ + ZA - (\mu P)GG^T(\mu P) + \Delta^TZ + Z\Delta + \mu(EF)^T(EF) + C^TJ^TQ^{-1} + Q^{-1}JC \\ & + Q^{-1}GG^TQ^{-1} + Q^{-1}JDG^TQ^{-1} + Q^{-1}G(JD)^TQ^{-1} + Q^{-1}JD(JD)^TQ^{-1}. \end{aligned}$$

The definition of Δ shows that $\Delta^TZ + Z\Delta + \mu(EF)^T(EF)$ is equal to

$$-(BF)^T(\mu P) - (\mu P)BF - \mu(H + EF)^T(H + EF) + \mu H^TH.$$

Therefore the (2, 2) block is given by

$$\begin{aligned} & A^TZ + ZA - (\mu P)GG^T(\mu P) - (BF)^T(\mu P) - (\mu P)BF - \mu(H + EF)^T(H + EF) \\ & + (JC)^TQ^{-1} + Q^{-1}(JC) + \mu H^TH + Q^{-1}(G + JD)(G + JD)^TQ^{-1}, \end{aligned}$$

which obviously equals

$$Q^{-1}R_JQ^{-1} - \mu R_F < 0. \quad \square$$

The modification Δ of the standard suboptimal compensator [4], [5] is due to the generality of F and J : We do not need to specify particularly chosen matrices F and J and can define a compensator based on certain solutions of the Riccati *inequalities*. The controller has a simple observer structure with a similar interpretation as discussed in [5]. In completely the same fashion we could also construct a compensator based on the minimal solutions of those Riccati equations that correspond to (3), (4) (which exist since $A + BF$ and $A + JC$ are stable). These solutions still satisfy (5) by minimality.

The question arises of how to test the various conditions of Theorem 1 in an algebraic way. The following results are available in the literature under combinations of the assumptions:

- (a) E has maximal column rank and D has maximal row rank.
- (b) $S(s)$ and $T(s)$ have no invariant zeros in \mathbb{C}^0 (see § 4).

If (b) holds, our characterization may be reformulated as follows [33]: There exist solutions $P \geq 0$, $Q \geq 0$ of the two *quadratic matrix inequalities*

$$\begin{aligned} & \begin{pmatrix} A^TP + PA + \mu PGG^TP + H^TH & PB + H^TE \\ B^TP + E^TH & E^TE \end{pmatrix} \geq 0 \text{ plus rank conditions,} \\ & \begin{pmatrix} AQ + QA^T + \mu QH^THQ + GG^T & QC^T + GD^T \\ CQ + DG^T & DD^T \end{pmatrix} \geq 0 \text{ plus rank conditions,} \end{aligned}$$

which satisfy

$$\rho(PQ) < \frac{1}{\mu}.$$

If both (a) and (b) hold true, the quadratic matrix inequalities can obviously be reduced (by taking the Schur complements with respect to the $(2, 2)$ blocks) to indefinite AREs and we arrive at the results of [4], [5]. If (a) fails to hold, the solutions of the quadratic matrix inequalities may as well be determined by solving two reduced-order indefinite AREs for subsystems of certain transformed versions of $S(s)$ and $T(s)$ [33]. A referee drew our attention to the paper [12] where algebraic tests for suboptimality are derived (in the frequency domain) if (a) and (b) do not necessarily hold true. However, these results are restricted to the one block problem, i.e., $T_2(s)$, $T_3(s)$ as appearing in § 2 must be square and invertible as rational matrices.

The aim of this paper is to remove both assumptions (a) and (b) for the general four block problem as described in § 2. We will generalize the suboptimality criteria we derived in [31] for the regular state-feedback problem ($C = I$, $D = 0$, E has maximal column rank) that are formulated in terms of the Riccati map

$$R(X, \mu) := AX + XA^T + XH^THX + \mu GG^T - (B - XH^TE)(E^TE)^+(B^T + E^THX)$$

and the associated map

$$A(X) := A + XH^TH - (B + XH^TE)(E^TE)^+E^TH,$$

defined on $\mathbb{S}^n \times \mathbb{R}$ and \mathbb{S}^n , respectively. Note that we performed in [31] a preliminary feedback with

$$(11) \quad F := -(E^TE)^+E^TH$$

(motivated by $E^T(H + EF) = 0$) and one easily verifies that $R(X, \mu)$ equals

$$(A + BF)X + X(A + BF)^T + X(H + EF)^T(H + EF)X + \mu GG^T - B(E^TE)^+B^T$$

and $A(X)$ is given by

$$(A + BF) + X(H + EF)^T(H + EF).$$

We already indicate the generalization to the singular problem by using the Moore-Penrose inverse: The two representations of these maps still coincide even if E does not have maximal column rank! The undetectable subspaces and the observability normal form appearing in [31] have their natural generalizations in the corresponding conditionally invariant subspaces and in a modification of Morse's canonical form, which are both thoroughly investigated in geometric control theory.

4. Transformation by restricted coordinate changes. For some arbitrary $(n + k) \times (n + m)$ system

$$S(s) = \begin{pmatrix} A - sI & B \\ H & E \end{pmatrix}$$

and any $\mathbb{C}^g \subset \mathbb{C}$ we denote by [1]

- $\sigma(S(s))$ the set of invariant zeros,
- $\mathcal{V}^g(S(s))$ the largest element under all subspaces $\mathcal{U} \subset \mathbb{R}^n$ for which there exists an F with $(A + BF)\mathcal{U} \subset \mathcal{U}$ and $\mathcal{U} \subset \ker(H + EF)$ such that the restriction of $A + BF$ to \mathcal{U} has only eigenvalues in \mathbb{C}^g .
- $\mathcal{S}^g(S(s))$ the smallest element under all subspaces $\mathcal{U} \subset \mathbb{R}^n$ for which there exists a J with $(A + JH)\mathcal{U} \subset \mathcal{U}$ and $\text{im}(B + JE) \subset \mathcal{U}$ such that $(A + JH)^T|_{\mathcal{U}^\perp}$ has only eigenvalues in \mathbb{C}^g .

We recall the important duality relation

$$\mathcal{S}_g(S(s)) = \mathcal{V}^g(S(s))^T^\perp.$$

For $\mathbb{C}^g = \mathbb{C}$, $\mathbb{C}^g = \mathbb{C}^-$, $\mathbb{C}^g = \mathbb{C}^0$, and $\mathbb{C}^g = \mathbb{C}^+$, we denote the corresponding spaces by replacing g with $*$, $-$, 0 and $+$, respectively. If we choose some \mathbb{C}^g with $\sigma(S(s)) \cap \mathbb{C}^g = \emptyset$, we define (independently of the choice of \mathbb{C}^g)

$$\mathcal{R}^*(S(s)) := \mathcal{V}^g(S(s)) \quad \text{and} \quad \mathcal{N}_*(S(s)) := \mathcal{S}_g(S(s)).$$

Motivated by the results in [31], we introduce for any $\lambda \in \mathbb{C}$ the *complex* subspace

$$\mathcal{V}^\lambda(S(s)) := \left\{ x \in \mathbb{C}^n \mid \exists u \in \mathbb{C}^m : S(\lambda) \begin{pmatrix} x \\ u \end{pmatrix} = 0 \right\}$$

as well as

$$\mathcal{S}_\lambda(S(s)) := \left\{ x \in \mathbb{C}^n \mid \exists u \in \mathbb{C}^{n+m} : \begin{pmatrix} x \\ 0 \end{pmatrix} = S(\lambda) u \right\}$$

and obtain the duality relation

$$\mathcal{S}_\lambda(S(s)) = \mathcal{V}^{\bar{\lambda}}(S(s))^T^\perp.$$

These various subspaces are displayed in the best way using the famous canonical form of Morse [22], which was given for not necessarily strictly proper systems in [1]. This canonical form is derived for the so-called full transformation group consisting of coordinate changes, “state-feedback” and “output-injection” transformations. The orbit of $S(s)$ under this transformation group can be easily described as

$$\left\{ LS(s)R \mid L = \begin{pmatrix} U & J \\ 0 & V \end{pmatrix}, R = \begin{pmatrix} U^{-1} & 0 \\ F & W \end{pmatrix}, U \in \mathbb{R}^{n \times n}, V, W \text{ square and nonsingular} \right\}.$$

The transformation properties $\mathcal{V}^g(LS(s)R) = U\mathcal{V}^g(S(s))$ and $\mathcal{S}_g(LS(s)R) = U\mathcal{S}_g(S(s))$ (where L and R are partitioned and denoted as above) are well known and we easily verify the same relations for the spaces $\mathcal{V}^\lambda(S(s))$ and $\mathcal{S}_\lambda(S(s))$. It will be instrumental to restrict this transformation group to pure *coordinate changes* ($F=0$ and $J=0$) or, since we work with the Euclidean norm in the output space, even to *restricted coordinate changes*: $F=0$, $J=0$; V and W orthogonal. Note that the set of coordinate changes and restricted coordinate changes still form a group.

The structure of the transformed system in the following result cannot be immediately extracted from [22], [1], not even for general coordinate transformations.

THEOREM 2. *The system*

$$S_F(s) := \begin{pmatrix} A + BF - sI & B \\ H + EF & E \end{pmatrix}$$

with F defined by (11) can be transformed by restricted coordinate changes to

$$\tilde{S}(s) = \begin{pmatrix} \tilde{A} - sI & \tilde{B} \\ \tilde{H} & \tilde{E} \end{pmatrix} := \begin{pmatrix} A_1 - sI & J_1 H_2 & 0 & \Sigma_1 \\ B_2 F_1 & A_2 - sI & B_2 & \Sigma_2 \\ H_1 & 0 & 0 & 0 \\ 0 & H_2 & 0 & 0 \\ 0 & 0 & 0 & \Sigma \end{pmatrix}$$

with the following properties:

- (a) Σ is symmetric and nonsingular;
- (b) $\begin{pmatrix} A_2 - \lambda I & B_2 \\ H_2 & 0 \end{pmatrix}$ has maximal row rank for all $\lambda \in \mathbb{C}$;

- (c) $\mathcal{S}_g(\tilde{S}(s)) = \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^n \mid x_1 \in \mathcal{S}_g \begin{pmatrix} A_1 - sI \\ H_1 \end{pmatrix} \right\} \text{ for any } \mathbb{C}^g;$
- (d) $\mathcal{S}_\lambda(\tilde{S}(s)) = \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{C}^n \mid x_1 \in \mathcal{S}_\lambda \begin{pmatrix} A_1 - sI \\ H_1 \end{pmatrix} \right\} \text{ for all } \lambda \in \mathbb{C}.$

The proof of this transformation result is based on an interesting solvability criterion for a linear equation which has its origin in the solution of the regulator problem (combine Korollar 7.2 with Satz 7.4 in [19]).

LEMMA 3. *Suppose that*

$$\begin{pmatrix} A - \lambda I & B \\ H & 0 \end{pmatrix}$$

of the size $(n+k) \times (n+m)$ has maximal row rank for all $\lambda \in \mathbb{C}$. Given any $M \in \mathbb{R}^{l \times l}$, $R \in \mathbb{R}^{n \times l}$, and $S \in \mathbb{R}^{k \times l}$ ($l \in \mathbb{N}$), there exist X and Y with

$$AX - XM + BY = R, \quad HX = S.$$

Proof of Proposition 2. The singular value decomposition yields orthogonal V, W and a nonsingular $\Sigma = \Sigma^T$ such that VEW equals $\begin{pmatrix} 0 & 0 \\ 0 & \Sigma \end{pmatrix}$. By the choice of F , we obtain

$$\begin{pmatrix} I & 0 \\ 0 & V \end{pmatrix} S_F(s) \begin{pmatrix} I & 0 \\ 0 & W \end{pmatrix} = \begin{pmatrix} \bar{A} - sI & \bar{B} & \bar{\Sigma} \\ \bar{H} & 0 & 0 \\ 0 & 0 & \Sigma \end{pmatrix}.$$

We now transform the strictly proper system

$$\tilde{S}(s) := \begin{pmatrix} \bar{A} - sI & \bar{B} \\ \bar{H} & 0 \end{pmatrix}$$

to Morse's canonical form [22]. If reversing the performed state-feedback and output-injection transformations and repartitioning suitably, the transformed system admits the shape

$$\left(\begin{array}{cc|c} \bar{A}_1 - sI & \bar{J}_1 \bar{H}_2 & 0 \\ \bar{A}_{21} & \bar{A}_2 - sI & \bar{B}_2 \\ \hline \bar{H}_1 & 0 & 0 \\ 0 & \bar{H}_2 & 0 \end{array} \right)$$

such that

$$\begin{pmatrix} \bar{A}_2 - \lambda I & \bar{B}_2 \\ \bar{H}_2 & 0 \end{pmatrix}$$

has maximal row rank for all $\lambda \in \mathbb{C}$. Since we can obviously reverse the coordinate change in the input space without violating these properties, this transformation is possible by general coordinate changes in the state space and output space; we denote the corresponding matrices by \bar{U} and \bar{V} . We clearly have

$$H\bar{U}^{-1} = \bar{V}^{-1} \begin{pmatrix} \bar{H}_1 & 0 \\ 0 & \bar{H}_2 \end{pmatrix}$$

and can decompose \bar{V}^{-1} as

$$\tilde{V}^T \begin{pmatrix} T_1 & 0 \\ T_2 & T_3 \end{pmatrix}$$

with some *orthogonal* \tilde{V} to obtain

$$\tilde{V}H\bar{U}^{-1} = \begin{pmatrix} T_1\bar{H}_1 & 0 \\ T_2\bar{H}_1 & T_3\bar{H}_2 \end{pmatrix}.$$

Therefore $\bar{S}(s)$ can be transformed by *restricted* coordinate changes to

$$(12) \quad \left(\begin{array}{cc|c} \bar{A}_1 - sI & J_1 H_2 & 0 \\ \bar{A}_{21} & \bar{A}_2 - sI & B_2 \\ \hline H_1 & 0 & 0 \\ H_{21} & H_2 & 0 \end{array} \right)$$

and, still,

$$\begin{pmatrix} \bar{A}_2 - \lambda I & B_2 \\ H_2 & 0 \end{pmatrix}$$

has maximal row rank for all $\lambda \in \mathbb{C}$.

At this point apply Lemma 3 to remove H_{21} and to shape \bar{A}_{21} . There exist matrices X and F_1 with

$$\bar{A}_2 X - X(\bar{A}_1 - J_1 H_{21}) + \bar{A}_{21} - B_2 F_1 = 0, \quad H_2 X + H_{21} = 0.$$

If we add the X -right multiple of the second column of (12) to the first one and then the $(-X)$ -left multiple of the first row to the second one (which is just a coordinate change in the state space), we cancel H_{21} and change \bar{A}_{21} to $B_2 F_1$ as desired. Since the block \bar{A}_2 is just transformed to $A_2 := \bar{A}_2 - X J_1 H_2$, condition (b) still holds.

If $S_2(s)$ denotes the system appearing in (b) and n_2 is the dimension of A_2 , the observations $\mathcal{S}_g(S_2(s)) = \mathbb{R}^{n_2}$ (for any \mathbb{C}^g) and $\mathcal{S}_\lambda(S_2(s)) = \mathbb{C}^{n_2}$ (for any $\lambda \in \mathbb{C}$) finish the proof. \square

The main reason for transforming to this special form is property (b) in Theorem 2. As it is known from the theory of almost disturbance decoupling, such a right-invertible system without any zeros has the following very appealing property [39], [36].

THEOREM 4. *If*

$$\begin{pmatrix} A - \lambda I & B \\ H & 0 \end{pmatrix}$$

has maximal row rank for all $\lambda \in \mathbb{C}$, there exists a sequence of feedback matrices F_j such that $A + B F_j$ is stable and

$$\lim_{j \rightarrow \infty} \|H(sI - A - B F_j)^{-1}\|_\infty = 0 \quad \text{as well as} \quad \lim_{j \rightarrow \infty} \int_0^\infty \|H e^{(A + B F_j)t}\|^2 dt = 0$$

hold true.

As noted in § 3, we will work with the Riccati map $R(\cdot, \cdot)$ and with $A(\cdot)$ even if E does not have maximal column rank. It is important to know how these maps behave under restricted coordinate changes. Let U denote the state-transformation matrix in $S(s) \rightarrow \tilde{S}(s)$, i.e., $U(A + BF)U^{-1} = \tilde{A}$. Then we have to transform G as

$$\tilde{G} = \begin{pmatrix} G_1 \\ G_2 \end{pmatrix} := UG.$$

We denote by $\tilde{R}(\cdot, \cdot)$ and $\tilde{A}(\cdot)$ the maps for the transformed data $\tilde{S}(s)$ and \tilde{G} , which are given by

$$\tilde{R}(X, \mu) = \tilde{A}X + X\tilde{A}^T + X\tilde{H}^T\tilde{H}X + \mu\tilde{G}\tilde{G}^T - \tilde{B}(\tilde{E}^T\tilde{E})^+\tilde{B} \quad \text{and} \quad \tilde{A}(X) = \tilde{A} + X\tilde{H}^T\tilde{H}.$$

The following result is verified by computation. The verification exhibits the reason for introducing the notion of restricted coordinate changes: We must make essential use of the orthogonality of *both* transformations in the input space and the output space.

LEMMA 5. *For any $X \in \mathbb{S}^n$ and $\mu \in \mathbb{R}$, the equations $UR(X, \mu)U^T = \tilde{R}(UXU^T, \mu)$ and $UA(X)U^{-1} = \tilde{A}(UXU^T)$ hold true.*

5. H_∞ -optimization by static state feedback. Our aim is to provide reasonable and verifiable characterizations for the set

$$\mathcal{P}_\mu := \{P > 0 \mid \exists F: (A + BF)^T P + P(A + BF) + \mu PGG^T P + (H + EF)^T(H + EF) < 0\}$$

to be nonempty if μ is positive.

The reader should recall the interpretation of the ARI (3) as explained in § 3, which shows that all the results of this section have immediate consequences for the general H_∞ -optimization problem by *static state feedback*. Let us denote its optimal value by

$$\mu_* := \sup \{ \|(H + EF)(sI - A - BF)^{-1}G\|_\infty^{-2} \mid \sigma(A + BF) \subset \mathbb{C}^- \}.$$

As an example, we mention the problem of *maximizing the complex stability radius by feedback*, which is just our H_∞ -optimization problem for $E = 0$ [13].

Remark. From Theorem 6 we extract the following well-known conclusion for the state-feedback H_∞ -problem: In the case where $C = I$ and $D = 0$, the optimal value μ_{opt} for *dynamic* controllers coincides with μ_* .

We define $S(s)$, G , $R(\cdot, \cdot)$, $A(\cdot)$ and the transformed objects $\tilde{S}(s)$, \tilde{G} , $\tilde{R}(\cdot, \cdot)$, $\tilde{A}(\cdot)$ as well as the state transformation matrix U as in § 4, where $\tilde{S}(s)$ has all the properties listed in Theorem 2. If any of the subspaces introduced in § 4 are defined with respect to $S(s)$, we drop the argument.

Our first suboptimality test is one of the central results of this paper.

THEOREM 6. *Suppose $\mu > 0$. Then $\mu < \mu_*$ if and only if the ARI*

$$(13) \quad A_1 X + X A_1^T + X H_1^T H_1 X + \mu G_1 G_1^T - (J_1 \quad \Sigma_1 \Sigma^{-1})(J_1 \quad \Sigma_1 \Sigma^{-1})^T < 0$$

has a positive definite solution.

Therefore $\mu < \mu_*$ can be tested by seeing whether or not a certain reduced-order algebraic Riccati inequality has a positive definite solution. The ARI is written in a suggestive way. If we introduce

$$(14) \quad B_1 := (J_1 \quad \Sigma_1 \Sigma^{-1}),$$

the ARI (13) has a positive definite solution if and only if $\mu > 0$ is suboptimal for the *regular* H_∞ -problem in which the underlying system is given by

$$(15) \quad S_1(s) := \begin{pmatrix} A_1 - sI & B_1 \\ H_1 & 0 \\ 0 & I \end{pmatrix}$$

and G_1 is the disturbance input matrix. This shows the only difference if comparing regular and singular problems: For regular problems, the suboptimality can be tested by looking at an ARI defined in terms of the original system matrices. For singular problems, we must transform the system and test the suboptimality for a certain (regular) subsystem. The reduction to the regular subsystem may be viewed as factoring out the “unproblematic” part of the system, which is constituted by the finite but assignable zeros and by the infinite zero structure (the \mathcal{S}_* -part of the system). The subsystem (15) is determined by the “relevant” part of the plant, which is determined

by the zeros of $S(s)$ (and in particular by those in \mathbb{C}^0 and \mathbb{C}^+) as well as the part of $S(s)$ “outside \mathcal{N}_* ,” which is related to the “observable part” of

$$\begin{pmatrix} A_1 - sI \\ H_1 \end{pmatrix}$$

and due to the nontriviality of H_1 [31, § 4].

Proof. By the very definition of restricted coordinate changes, we can work without restriction with the transformed data $\tilde{S}(s)$, \tilde{G} .

Suppose $\mu < \mu_*$ and choose some F and $X > 0$ with

$$(16) \quad (\tilde{A} + \tilde{B}F)X + X(\tilde{A} + \tilde{B}F)^T + X(\tilde{H} + \tilde{E}F)^T(\tilde{H} + \tilde{E}F)X + \mu\tilde{G}\tilde{G}^T < 0.$$

We now partition

$$X = \begin{pmatrix} X_1 & X_{12} \\ X_{12}^T & X_2 \end{pmatrix}$$

according to \tilde{A} and FX as $\begin{pmatrix} * & * \\ * & * \end{pmatrix}$ according to the column partition of \tilde{B} and the row partition of \tilde{A} . Then we compute the (1, 1) block of the ARI (16) to obtain (recalling $\Sigma = \Sigma^T$)

$$\begin{aligned} & A_1X_1 + X_1A_1^T + J_1H_2X_{12}^T + (J_1H_2X_{12}^T)^T + \Sigma_1\tilde{F} + (\Sigma_1\tilde{F})^T \\ & + X_1H_1^TH_1X_1 + X_{12}H_2^TH_2X_{12}^T + \tilde{F}^T\Sigma^2\tilde{F} + \mu G_1G_1^T < 0. \end{aligned}$$

An obvious completion of the squares delivers

$$\begin{aligned} & A_1X_1 + X_1A_1^T + X_1H_1^TH_1X_1 + \mu G_1G_1^T - J_1J_1^T - \Sigma_1\Sigma^{-2}\Sigma_1 \\ & + (J_1 + X_{12}H_2^T)(J_1 + X_{12}H_2^T)^T + (\Sigma^{-2}\Sigma_1^T + \tilde{F})^T\Sigma^2(\Sigma^{-2}\Sigma_1^T + \tilde{F}) < 0. \end{aligned}$$

This already proves the necessity part.

For the proof of sufficiency it is enough to construct some \tilde{F} with

$$(17) \quad \sigma(\tilde{A} + \tilde{B}\tilde{F}) \subset \mathbb{C}^-$$

and such that the H_∞ -norm of

$$(18) \quad (\tilde{H} + \tilde{E}\tilde{F})(sI - \tilde{A} - \tilde{B}\tilde{F})^{-1}\tilde{G}$$

is smaller than $1/\sqrt{\mu}$.

Now suppose that the ARI (13) has the positive definite solution X_1 . Then there exists some F (e.g., $F := -B_1^TX_1^{-1}$) [31] such that

$$A_1 + B_1F$$

is stable and the inequality

$$(19) \quad \left\| \begin{pmatrix} H_1 \\ F \end{pmatrix} H(s) G_1 \right\|_\infty < \frac{1}{\sqrt{\mu}}$$

holds for $H(s) := (A_1 + B_1F - sI)^{-1}$. Let us partition $F = \begin{pmatrix} F_s \\ F_z \end{pmatrix}$ according to the column partitioning of B_1 in (14).

Exploiting the properties of the structure at infinity, it is possible to approximate the transfer matrix in (19) by (18) in the H_∞ -norm, if \tilde{F} is a suitably chosen feedback

matrix with (17). If the error is small enough, the proof is finished. An obvious feedback transformation of $\tilde{S}(s)$ leads to

$$(20) \quad \begin{pmatrix} A_1 + \Sigma_1 \Sigma^{-1} F_\Sigma - sI & J_1 H_2 & 0 & \Sigma_1 & G_1 \\ 0 & A_2 - sI & B_2 & \Sigma_2 & G_2 \\ H_1 & 0 & 0 & 0 & 0 \\ 0 & H_2 & 0 & 0 & 0 \\ F_\Sigma & 0 & 0 & \Sigma & 0 \end{pmatrix}.$$

According to Lemma 3, there exist X and Y with

$$A_2 X - X(A_1 + B_1 F) - B_2 Y = 0 \quad \text{and} \quad H_2 X = F_j.$$

We now add the X -right multiple of the second column of (20) to the first one and the $(-X)$ -left multiple of the first row of the resulting pencil to its second row (which amounts to a coordinate change for (20) in the state-space). After a suitable feedback to eliminate the $(2, 1)$ block, we obtain

$$\begin{pmatrix} A_1 + B_1 F - sI & J_1 H_2 & 0 & \Sigma_1 & G_1 \\ 0 & \tilde{A}_2 - sI & B_2 & \tilde{\Sigma}_2 & \tilde{G}_2 \\ H_1 & 0 & 0 & 0 & 0 \\ F_j & H_2 & 0 & 0 & 0 \\ F_\Sigma & 0 & 0 & \Sigma & 0 \end{pmatrix}$$

with $\tilde{A}_2 := A_2 - XJ_1 H_2$, $\tilde{\Sigma}_2 := \Sigma_2 - X\Sigma_1$ and $\tilde{G}_2 := G_2 - XG_1$. Therefore

$$\begin{pmatrix} \tilde{A}_2 - \lambda I & B_2 \\ H_2 & 0 \end{pmatrix}$$

still has maximal row rank for all $\lambda \in \mathbb{C}$. By Theorem 4, we can hence construct a sequence $F_2(j)$ such that $\tilde{A}_2 + B_2 F_2(j)$ is stable and

$$(21) \quad \lim_{j \rightarrow \infty} \|H_2 H_j(s)\|_\infty^2 = 0$$

holds for $H_j(s) := (\tilde{A}_2 + B_2 F_2(j) - sI)^{-1}$. A last feedback finally leads to the closed-loop system

$$\left(\begin{array}{cc|c} A_1 + B_1 F - sI & J_1 H_2 & G_1 \\ 0 & \tilde{A}_2 + B_2 F_2(j) - sI & \tilde{G}_2 \\ \hline H_1 & 0 & 0 \\ F_j & H_2 & 0 \\ F_\Sigma & 0 & 0 \end{array} \right).$$

This system is internally stable. Moreover, its transfer matrix equals

$$(22) \quad \begin{pmatrix} H_1 & 0 \\ F_j & H_2 \\ F_\Sigma & 0 \end{pmatrix} \begin{pmatrix} H(s) & -H(s)J_1 H_2 H_j(s) \\ 0 & H_j(s) \end{pmatrix} \begin{pmatrix} G_1 \\ \tilde{G}_2 \end{pmatrix}$$

by the formula for the inverse of a block matrix. Hence we can infer from (19) and (21) that the H_∞ -norm of (22) is less than $1/\sqrt{\mu}$ for some j which is sufficiently large. \square

The proof contains an explicit construction of stabilizing suboptimal feedback matrices which yield a closed-loop system with a block triangular structure. Since we make no *special* choices for the μ -suboptimal feedbacks for the regular H_∞ -subproblem (called F) and the sequence of feedbacks on the infinite zero structure used for the approximation (denoted as $F_2(j)$), it is possible to take into account additional freedom (e.g., pole placement requirements) or simplifications for the construction of these matrices [31], [36].

If we combine Theorem 6 with the results in [31] we conclude:

The optimal value μ_* of the general state-feedback H_∞ -problem for $S(s)$ and G can be computed by a quadratically convergent algorithm. The infinite zeros of $S(s)$ only cause additional *algebraic* effort for transforming $S(s)$ into $\tilde{S}(s)$.

Another technique to construct suboptimal static feedback matrices is to perturb H and E such that the perturbation of E has maximal column rank. Since we want to allow for rather general perturbation schemes, we introduce the following notion.

DEFINITION 7. The family $(0, \infty) \ni \varepsilon \rightarrow (H_\varepsilon \ E_\varepsilon) \in \mathbb{R}^{(k+p) \times (n+m)}$, $p \in \mathbb{N}_0$, is called an *admissible perturbation* of $(H \ E) \in \mathbb{R}^{k \times (n+m)}$ if the following conditions are satisfied:

- (a) $\lim_{\varepsilon \rightarrow 0} (H_\varepsilon \ E_\varepsilon) = \begin{pmatrix} H & E \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{(k+p) \times (n+m)}$;
- (b) $(H \ E)^T (H \ E) \leq (H_\varepsilon \ E_\varepsilon)^T (H_\varepsilon \ E_\varepsilon)$ for $\varepsilon > 0$;
- (c) E_ε has maximal column rank for $\varepsilon > 0$.

THEOREM 8. Suppose that $P > 0$ satisfies

$$(23) \quad A^T P + P A + H_\varepsilon^T H_\varepsilon + \mu P G G^T P - (P B + H_\varepsilon^T E_\varepsilon)(E_\varepsilon^T E_\varepsilon)^{-1}(E_\varepsilon^T H_\varepsilon + B^T P) < 0$$

for some $\varepsilon > 0$. If defining F by

$$F := -(E_\varepsilon^T E_\varepsilon)^{-1}(E_\varepsilon^T H_\varepsilon + B^T P),$$

P satisfies the ARI (3) and is thus an element of \mathcal{P}_μ . If P is contained in \mathcal{P}_μ , there exists an $\varepsilon > 0$ such that P solves (23).

Proof. According to the different possible representations of the Riccati map as presented in § 4, the ARI (23) can be rearranged to

$$(24) \quad (A + B F)^T P + P(A + B F) + \mu P G G^T P + (H_\varepsilon + E_\varepsilon F)^T (H_\varepsilon + E_\varepsilon F) < 0.$$

The inequality (b) implies $(H + E F)^T (H + E F) \leq (H_\varepsilon + E_\varepsilon F)^T (H_\varepsilon + E_\varepsilon F)$, which yields $P \in \mathcal{P}_\mu$. The convergence property (a) shows that there exists for $P \in \mathcal{P}_\mu$ some F and some $\varepsilon > 0$ for which (24) holds. Since E_ε has maximal column rank by (c), a standard completion of the squares argument leads to the ARI (23). \square

The union of all positive definite solutions of (23) over all $\varepsilon > 0$ therefore coincides with \mathcal{P}_μ . This result allows us to construct a suboptimal feedback matrix by finding a suitable $\varepsilon > 0$, in fact, directly in terms of the system data without any preliminary transformation. In practice, however, the choice of $\varepsilon > 0$ remains as a problem.

From a computational point of view, we arrive at a satisfactory solution of the state-feedback H_∞ -problem. However, we wish as well to have an algebraic characterization (without referring to perturbations) of suboptimality for which no preliminary transformation of the plant is required and which is directly formulated in terms of the original data $S(s)$ and G . Indeed it is possible to generalize the criterion we derived in [31]. We first define the set $\mathcal{T}(\mu)$ (apart from an additional invariance requirement) literally as in [31, § 4].

DEFINITION 9. For any $\mu \in \mathbb{R}$, $\mathcal{T}(\mu)$ is the set of all matrices $Z \in \mathbb{S}^n$ such that

- \mathcal{S}_* is $A(Z)$ -invariant,
- the restriction of $A(Z)^T$ to \mathcal{S}_+^\perp has all its eigenvalues in \mathbb{C}^+ ,
- $x^T R(Z, \mu) y$ vanishes for $x \in \mathcal{S}_+^\perp$ and $y \in \mathcal{S}_+^\perp + \mathcal{S}_0^\perp$.

We stress that $A(Z)\mathcal{S}_* \subset \mathcal{S}_*$ implies $A(Z)^T \mathcal{S}_+^\perp \subset \mathcal{S}_+^\perp$ such that the definition makes sense. Now we can formulate an abstract but algebraically verifiable suboptimality test in terms of the original data matrices.

THEOREM 10. Suppose $\mu > 0$. Then $\mu < \mu_*$ if and only if there exists some $Z \in \mathcal{T}(\mu)$ such that

$$(25) \quad Z \text{ is positive on } \mathcal{S}_+^\perp \text{ and } R(Z, \mu) \text{ is negative on } \bigcup_{\lambda \in \mathbb{C}^0} \mathcal{S}_\lambda^\perp \setminus \mathcal{S}_+^\perp.$$

How do the matrices in $\mathcal{T}(\mu)$ look and how should these conditions be understood? The answer is simple and is motivated by Theorem 6. According to Definition 9, we introduce the set $\mathcal{T}_1(\mu)$ for the Riccati map

$$R_1(X, \mu) := A_1 X + X A_1^T + X H_1^T H_1 X + \mu G_1 G_1^T - B_1 B_1^T,$$

where the underlying system is $S_1(s)$ as defined in (15). By

$$(26) \quad \mathcal{S}_g(S_1(s)) = \mathcal{S}_g \begin{pmatrix} A_1 - sI \\ H_1 \end{pmatrix}$$

for any \mathbb{C}^g or $g = \lambda \in \mathbb{C}$, this is just the linear manifold we discussed intensively in [31]. The following result displays the simple relation of $\mathcal{T}(\mu)$ and $\mathcal{T}_1(\mu)$. Moreover, the combination of Theorem 6 and [31, Thm. 10] with the properties of the quadratic forms appearing in the Lemma 11 proves our abstract suboptimality criterion.

LEMMA 11. It holds that

$$\mathcal{T}(\mu) = U^{-1} \left\{ \begin{pmatrix} Z_1 & Z_{12} \\ Z_{12}^T & Z_2 \end{pmatrix} \in \mathbb{S}^n \mid Z_1 \in \mathcal{T}_1(\mu), Z_{12} \text{ satisfies } J_1 + Z_{12} H_2^T = 0 \right\} U^{-T}.$$

Moreover, if choosing $Z \in \mathcal{T}(\mu)$ and $Z_1 \in \mathcal{T}_1(\mu)$, then

$$\begin{aligned} \{x^* Z x \mid x \in \mathcal{S}_+^\perp\} &= \{x^* Z_1 x \mid x \in \mathcal{S}_+(S_1(s))^\perp\}, \\ \{x^* R(Z, \mu) x \mid x \in \mathcal{S}_\lambda^\perp\} &= \{x^* R_1(Z_1, \mu) x \mid x \in \mathcal{S}_\lambda(S_1(s))^\perp\} \end{aligned}$$

hold for any $\lambda \in \mathbb{C}^0$.

Proof. Lemma 5 and the transformation properties of $\mathcal{S}_g(\cdot)$ for $g = +, 0, \lambda \in \mathbb{C}$ show that we can prove the lemma without restriction for the system $\tilde{S}(s)$ and \tilde{G} . If partitioning any $Z \in \mathbb{S}^n$, $\tilde{A}(Z)$ and $\tilde{R}(Z, \mu)$ as \tilde{A} , we easily compute

$$\tilde{A}(Z) = \begin{pmatrix} A_1 + Z_1 H_1^T H_1 & J_1 H_2 + Z_{12} H_2^T H_2 \\ * & * \end{pmatrix}$$

and the blocks $\tilde{R}_1(Z, \mu)$, $\tilde{R}_{12}(Z, \mu)$, and $\tilde{R}_2(Z, \mu)$ of $\tilde{R}(Z, \mu)$, which are given by

$$\begin{aligned} &A_1 Z_1 + Z_1 A_1^T + Z_1 H_1^T H_1 Z_1 + \mu G_1 G_1^T - \Sigma_1 \Sigma^{-2} \Sigma_1^T - J_1 J_1^T + (J_1 + Z_{12} H_2^T)(J_1 + Z_{12} H_2^T)^T, \\ &(A_1 + Z_1 H_1^T H_1) Z_{12} + Z_{12} A_2^T + Z_1 F_1^T B_2^T + (J_1 + Z_{12} H_2^T) H_2 Z_2 + \mu G_1 G_2^T - \Sigma_1 \Sigma^{-2} \Sigma_2^T, \\ &A_2 Z_2 + Z_2 A_2^T + Z_2 H_2^T H_2 Z_2 - \Sigma_2 \Sigma^{-2} \Sigma_2^T + \mu G_2 G_2^T + Z_{12}^T H_1^T H_1 Z_{12} + B_2 F_1 Z_{12} + Z_{12}^T F_1^T B_2^T, \end{aligned}$$

respectively. The combination of Theorem 2 with (26) implies

$$\mathcal{S}_g^\perp(\tilde{S}(s)) = \left\{ x = \begin{pmatrix} x_1 \\ 0 \end{pmatrix} \mid x_1 \in \mathcal{S}_g(S_1(s))^\perp \right\}$$

for all \mathbb{C}^g or $g = \lambda \in \mathbb{C}$.

We infer that $\tilde{A}(Z)$ leaves $\mathcal{S}_*(\tilde{S}(s))$ invariant if and only if $J_1 H_2 + Z_{12} H_2^T H_2 = 0$, or, since H_2 has maximal row rank, $J_1 + Z_{12} H_2^T = 0$.

Now suppose that $J_1 + Z_{12}H_2^T$ vanishes. Then $\tilde{A}(Z)^T$ is antistable on $\mathcal{S}_+(\tilde{S}(s))^\perp$ if and only if $(A_1 + Z_1H_1^TH_1)^T$ is antistable on $\mathcal{S}_+(S_1(s))^\perp$. Moreover we obviously have $\tilde{R}_1(Z, \mu) = R_1(Z_1, \mu)$. If we choose $x = (x_1^* 0)^* \in \mathcal{S}_+^\perp(\tilde{S}(s))$ and $y = (y_1^* 0) \in \mathcal{S}_+^\perp(\tilde{S}(s)) + \mathcal{S}_0^\perp(\tilde{S}(s))$, we infer $x^TR(Z, \mu)y = x_1^TR_1(Z_1, \mu)y_1$. This already proves the relation of $\mathcal{T}(\mu)$ and $\mathcal{T}_1(\mu)$.

Now take $x = (x_1^* 0)^* \in \mathcal{S}_+^\perp(\tilde{S}(s))$ and $y = (y_1^* 0)^* \in \mathcal{S}_\lambda^\perp(\tilde{S}(s))$ for some $\lambda \in \mathbb{C}^0$. We infer $x^*Zx = x_1^*Z_1x_1$ and $y^TR(Z, \mu)y = y_1^TR_1(Z_1, \mu)y_1$ and just have to note that neither $x_1^*Z_1x_1$ nor $y_1^TR_1(Z_1, \mu)y_1$ depend on the particular choice of $Z_1 \in \mathcal{T}_1(\mu)$ to finish the proof. \square

Remark. Lemma 11 shows that $\mathcal{T}(\mu)$, if nonempty, is a linear manifold in \mathbb{S}^n . Moreover, the positivity/negativity conditions (25) hold for *one* $Z \in \mathcal{T}(\mu)$ if and only if they hold for *all* $Z \in \mathcal{T}(\mu)$. Therefore it suffices to choose one element $Z \in \mathcal{T}(\mu)$ and to verify (25) for this particular Z . This can be done explicitly as explained in [31].

We should recall the following intuitive interpretation of $\mathcal{T}(\mu)$: $\mathcal{T}(\mu)$ is nonempty if and only if a certain reduced-order Riccati equation has a (unique) antistabilizing solution. A parametrization of $\mathcal{T}(\mu)$ is computed by solving this ARE and two (always solvable) linear equations. Positivity of $\mathcal{T}(\mu)$ on \mathcal{S}_+^\perp is related to the positive definiteness of the ARE solution. The negativity of $R(\mathcal{T}(\mu), \mu)$ on $\mathcal{S}_\lambda^\perp \setminus \mathcal{S}_+^\perp$ for all $\lambda \in \mathbb{C}^0$ is related to the solvability of a reduced-order Lyapunov inequality.

Let us introduce the following critical parameters:

$$\begin{aligned} \mu_{\max}: \quad & \mu < \mu_{\max} \Leftrightarrow \mathcal{T}(\mu) \neq \emptyset, \\ \mu_{\text{pos}}: \quad & \mu < \mu_{\text{pos}} \Leftrightarrow \mu < \mu_{\max}, \mathcal{T}(\mu) \text{ is positive on } \mathcal{S}_+^\perp, \\ \mu_{\text{neg}}: \quad & \mu < \mu_{\text{neg}} \Leftrightarrow \mu < \mu_{\max}, R(\mathcal{T}(\mu), \mu) \text{ is negative on } \bigcup_{\lambda \in \mathbb{C}^0} \mathcal{S}_\lambda^\perp \setminus \mathcal{S}_+^\perp, \end{aligned}$$

which obviously lead to

$$\mu_* = \min \{\mu_{\text{pos}}, \mu_{\text{neg}}\}.$$

We can again refer to [31] for the fast computation of all these parameters since they coincide with those for $T_1(\mu)$. Moreover, μ_* coincides with μ_{pos} if $S(s)$ has no invariant zeros on the imaginary axis. Due to the \mathbb{C}^0 -zeros of $S(s)$, the optimal value may jump from μ_{pos} to $\min \{\mu_{\text{pos}}, \mu_{\text{neg}}\}$. In fact, μ_{neg} has its own system theoretic significance.

Remark. Suppose that $(A - sI \ B)$ is stabilizable with respect to \mathbb{C}^0 . Along the lines of our approach to the H_∞ -problem and exploiting [31, Thm. 3(a)], we can prove for any $\mu > 0$ the following equivalences:

$$\begin{aligned} \exists F: \quad & \sigma(A + BF) \cap \mathbb{C}^0 = \emptyset, \quad \mu < \|(H + EF)(sI - A - BF)^{-1}G\|_\infty^{-2} \\ & \Leftrightarrow \\ \exists P \in \mathbb{S}^n: \quad & (A + BF)^TP + P(A + BF) + \mu PGG^TP + (H + EF)^T(H + EF) < 0 \\ & \Leftrightarrow \\ & \mu < \mu_{\text{neg}}. \end{aligned}$$

This implies that μ_{neg} is the optimal value of the L_∞ -optimization problem

$$\sup \{ \|(H + EF)(sI - A - BF)^{-1}G\|_\infty^{-2} \mid \sigma(A + BF) \cap \mathbb{C}^0 = \emptyset \}.$$

Let us conclude this section with an algebraic characterization of $\mu_{\max} = \infty$, $\mu_{\text{pos}} = \infty$, and $\mu_{\text{neg}} = \infty$. By [31, Thm. 13] the following result no longer needs proof.

THEOREM 12. (a) $\mu_{\max} = \infty$ if and only if $\text{im}(G) \subset \mathcal{N}_*(S(s))$. In the case of $\mu_{\max} = \infty$, the parameters μ_{pos} and μ_{neg} can be computed by solving Hermitian eigenvalue problems.

- (b) $\mu_{\text{pos}} = \infty$ if and only if $\text{im}(G) \subset \mathcal{S}_+$.
 (c) $\mu_{\text{neg}} = \infty$ if and only if $\text{im}(G) \subset \bigcap_{\lambda \in \mathbb{C}^0} \mathcal{S}_\lambda$.

6. The solution of the general H_∞ -optimization problem. We have discussed in § 5 how to test algebraically whether or not \mathcal{P}_μ is nonempty. If we introduce

$$\mathcal{Q}_\nu := \{Q > 0 \mid \exists J: (A + JC)Q + Q(A + JC)^T + \nu QH^T H Q + (G + JD)(G + JD)^T < 0\},$$

it is possible by dualization to check $\mathcal{Q}_\nu \neq \emptyset$ in the same way. We have even shown how to compute the critical parameters μ_* and ν_* such that $\mathcal{P}_\mu, \mathcal{Q}_\nu$ are nonempty if and only if $\mu < \mu_*, \nu < \nu_*$. Therefore it remains to verify the coupling condition (5) for $\mu < \min\{\mu_*, \nu_*\}$. If we tried to test this condition directly, we would be confronted with the difficulty of finding *suitable* $P \in \mathcal{P}_\mu$ and $Q \in \mathcal{Q}_\mu$, but is it not clear how to make a concrete choice. At this point the introduction of lower limit as discussed in [31] for general Riccati inequalities proves to be very fruitful. Let $P_1(\mu)$ denote the strict lower limit point of the set of inverses of all positive definite solutions of (13). We stress that it is simple to compute $P_1(\mu)$ by solving one well-defined reduced-order Riccati equation [31, Thm. 9].

THEOREM 13. *For $\mu < \mu_*$, the set \mathcal{P}_μ has a strict lower limit point $P(\mu)$, which is given by*

$$P(\mu) = U^T \begin{pmatrix} P_1(\mu) & 0 \\ 0 & 0 \end{pmatrix} U.$$

Dually, the set \mathcal{Q}_ν has a *computable* strict lower limit point $Q(\nu)$ for $\nu < \nu_*$. This solves our problem: Fix $\mu < \min\{\mu_*, \nu_*\}$. Suppose that $P \in \mathcal{P}_\mu$ and $Q \in \mathcal{Q}_\mu$ satisfy (5). By $P(\mu) < P$ and $Q(\mu) < Q$, we obtain

$$(27) \quad \rho(P(\mu)Q(\mu)) < \frac{1}{\mu}.$$

Now assume (27). There exist sequences $P_j \in \mathcal{P}_\mu$ and $Q_j \in \mathcal{Q}_\mu$ with $P_j \rightarrow P(\mu)$ and $Q_j \rightarrow Q(\mu)$ for $j \rightarrow \infty$. By (27), there exists a sufficiently large j with $\rho(P_j Q_j) < 1/\mu$. We arrive at the central result of this paper.

THEOREM 14. *The positive parameter μ satisfies $\mu < \mu_{\text{opt}}$ if and only if*

$$\mu < \mu_*, \quad \mu < \nu_*, \quad \text{and} \quad \rho(P(\mu)Q(\mu)) < \frac{1}{\mu}.$$

As shown in § 5, the inequalities $\mu < \mu_*$ and $\mu < \nu_*$ can be tested algebraically. Moreover, we recall the detailed investigation of the influence of the various zeros of $S(s)$ and $T(s)$ on both values μ_* and ν_* in [31, § 4]. From the explicit formula for $P(\mu)$ and, dually, for $Q(\mu)$ it is obvious that these matrices are *not influenced by the \mathbb{C}^0 -zeros* of $S(s)$ and $T(s)$, which may be expressed as follows:

The \mathbb{C}^0 -zeros of $S(s)$ and $T(s)$ do not cause additional coupling constraints.

It remains to prove Theorem 13, which is again based on the nice properties of the subsystem of $\tilde{S}(s)$ appearing in Theorem 2(b). In particular we exhibit in Proposition 15 the striking flexibility of the Riccati map defined for a right invertible system without any zeros. Its proof is given in the Appendix.

PROPOSITION 15. *Let*

$$S(\lambda) = \begin{pmatrix} A - \lambda I & B \\ H & 0 \end{pmatrix}$$

have maximal row rank for all $\lambda \in \mathbb{C}$. Then for all $X_0 \in \mathbb{S}^n$ and $R_0 \in \mathbb{S}^n$ there exist an $\varepsilon > 0$ and an $X \in \mathbb{S}^n$ with

$$X > X_0 \quad \text{and} \quad AX + XA^T + XH^T HX - \frac{BB^T}{\varepsilon^2} < R_0.$$

Proof of Theorem 13. We can again assume without restriction that

$$S(s) = \tilde{S}(s) \quad \text{with} \quad F_1 = 0 \quad \text{and} \quad G = \tilde{G}.$$

Choose $P \in \mathcal{P}_\mu$. Take any F such that $X := P^{-1}$ satisfies (16). We proved that the $(1, 1)$ block X_1 of X solves the ARI (13) and we infer by the definition of $P_1(\mu)$ the inequality $P_1(\mu) < X_1^{-1}$. Lemma 14 in [31] shows one part of the claim:

$$P(\mu) := \begin{pmatrix} P_1(\mu) & 0 \\ 0 & 0 \end{pmatrix} < P.$$

Let us now define the admissible perturbation

$$\tilde{H}_\varepsilon := \begin{pmatrix} H_1 & 0 \\ 0 & H_2 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad \tilde{E}_\varepsilon := \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & \Sigma \\ \varepsilon I & 0 \end{pmatrix}$$

and, for notational simplicity, the Riccati map

$$\tilde{R}_\varepsilon(X, \mu) := \tilde{A}X + X\tilde{A}^T + X\tilde{H}_\varepsilon^T \tilde{H}_\varepsilon X + \mu \tilde{G}\tilde{G}^T - \tilde{B}(\tilde{E}_\varepsilon^T \tilde{E}_\varepsilon)^{-1} \tilde{B}^T.$$

According to Theorem 8, it suffices to construct sequences $\varepsilon_j > 0$ and $X(j) > 0$ which satisfy $\tilde{R}_{\varepsilon_j}(X(j), \mu) < 0$ and $X(j)^{-1} \rightarrow P(\mu)$ for $j \rightarrow \infty$. We exploit the obvious relation

$$\tilde{R}_\varepsilon(X, \mu) = \tilde{R}(X, \mu) + \begin{pmatrix} 0 & 0 \\ 0 & -(1/\varepsilon^2)B_2B_2^T \end{pmatrix}$$

and note that we already computed the blocks of $\tilde{R}(X, \mu)$ in the proof of Lemma 11. By the definition of $P_1(\mu)$, we can find a sequence $X_1(j) > 0$ of solutions of the ARI (13) with $X_1(j)^{-1} \rightarrow P_1(\mu)$ for $j \rightarrow \infty$. Moreover, we define $X_{12}(j) \equiv X_{12}$, where X_{12} satisfies $J_1 + X_{12}H_2^T = 0$. If we look at the structure of $\tilde{R}_2(\cdot, \mu) - (1/\varepsilon^2)B_2B_2^T$, we can deduce from Proposition 15 the existence of sequences $\varepsilon_j > 0$ and $X_2(j)$ such that

$$X(j) := \begin{pmatrix} X_1(j) & X_{12} \\ X_{12}^T & X_2(j) \end{pmatrix}$$

is positive definite, $X_2(j)^{-1}$ converges to 0 for $j \rightarrow \infty$, and $\tilde{R}_{\varepsilon_j}(X(j), \mu)$ is negative definite. Again by [31, Lemma 14], we infer $X(j)^{-1} \rightarrow P(\mu)$ for $j \rightarrow \infty$. \square

7. The computation of the optimal value. We have intensively discussed [30], [31] how to compute the critical parameters μ_{pos} , μ_{neg} , and μ_{max} for $(S(s) \ G)$ and those for $(T(s)^T \ H^T)$ denoted as ν_{pos} , ν_{neg} , and ν_{max} .

In § 6 we introduced $P(\cdot)$ on $(-\infty, \mu_*)$ and $Q(\cdot)$ on $(-\infty, \nu_*)$ and proved that we only have to find in addition

$$\mu_{\text{cou}} := \sup \left\{ \mu < \min \{ \mu_*, \nu_* \} \mid \rho(P(\mu)Q(\mu)) < \frac{1}{\mu} \right\}$$

to determine μ_{opt} .

If we combine Theorem 13 with [31, Thm. 9], there exist nonsingular constant matrices S, T with

$$P(\mu) = S^T \begin{pmatrix} X(\mu)^{-1} & 0 \\ 0 & 0 \end{pmatrix} S, \quad X(\mu) > 0, \quad \text{for } \mu < \mu_*$$

and

$$Q(\nu) = T^T \begin{pmatrix} Z(\nu)^{-1} & 0 \\ 0 & 0 \end{pmatrix} T, \quad Z(\nu) > 0, \quad \text{for } \nu < \nu_*.$$

Let us denote the left upper block of ST^T (in an obvious partition) as V . Then we extract from [30, § 7] for $\mu \in (-\infty, \min \{\mu_*, \nu_*\})$ the equivalence

$$\rho(P(\mu)Q(\mu)) < \frac{1}{\mu} \Leftrightarrow \mu V^T Z(\mu)^{-1} V < X(\mu).$$

If we define the analytic function

$$(-\infty, \min \{\mu_*, \nu_*\}) \ni \mu \rightarrow F(\mu) := X(\mu) - \mu V^T Z(\mu)^{-1} V,$$

we obtain

$$\mu_{\text{cou}} = \sup \{ \mu \mid 0 < \mu < \min \{\mu_*, \nu_*\}, F(\mu) > 0 \}.$$

Since $F(0)$ is positive definite and $F'(\mu)$ as well as $F''(\mu)$ are both negative semidefinite for $\mu < \min \{\mu_*, \nu_*\}$ (because both $X(\cdot)$ and $Y(\cdot)$ have these properties), we can literally apply to $F(\cdot)$ the algorithm formulated in [30, § 7] to compute μ_{cou} and then μ_{opt} . As explained in [31, § 5], we can combine the computation of μ_{cou} with that of μ_*, ν_* , which implies that there exists a quadratically convergent algorithm to determine μ_{opt} .

If one of the values μ_{max} or ν_{max} is infinite, $\mu_{\text{pos}}, \mu_{\text{neg}}$ or $\nu_{\text{pos}}, \nu_{\text{neg}}$ can be found by solving linear equations and Hermitian eigenvalue problems [30], [31], and it may happen, depending in particular on V , that $F(\cdot)$ also has simple structural properties which allow the explicit computation of μ_{opt} . If both values are infinite, this can be assured and we summarize the consequences in the following result.

THEOREM 16. $\mu_{\text{max}} = \infty$ if and only if $\text{im}(G) \subset \mathcal{N}_*(S(s))$ and, dually, $\nu_{\text{max}} = \infty$ if and only if $\mathcal{R}^*(T(s)) \subset \ker(H)$. If both values are infinite, there exists a computable square polynomial matrix $P_{\text{cou}}(\mu)$ such that the optimal value is given by

$$\min \{ \mu_*, \nu_* \}$$

if $P_{\text{cou}}(\mu)$ has no zeros in $(0, \min \{ \mu_*, \nu_* \})$ or, otherwise, by

$$\mu_{\text{opt}} = \min \{ \mu \in (0, \min \{ \mu_*, \nu_* \}) \mid \det(P_{\text{cou}}(\mu)) = 0 \}.$$

Proof. We only have to consider the characterization of μ_{opt} in the case of $\mu_{\text{max}} = \infty$ and $\nu_{\text{max}} = \infty$. Then $X(\cdot)$ and $Y(\cdot)$ are both affine and, therefore, $F(\cdot)$ is a real rational function without poles in $(-\infty, \min \{ \mu_*, \nu_* \})$. Let $d(\cdot)$ denote the least common multiple of the denominators of all the elements of $F(\cdot)$ and define the polynomial matrix $P_{\text{cou}}(\mu) := d(\mu)F(\mu)$. We obtain for any $\mu \in (-\infty, \min \{ \mu_*, \nu_* \})$:

$$\det(F(\mu)) = 0 \Leftrightarrow \det(P_{\text{cou}}(\mu)) = 0.$$

The distinction between the following cases finishes the proof. $P_{\text{cou}}(\mu)$ has no zeros in $(0, \min \{ \mu_*, \nu_* \})$: Then $\det(F(\mu))$ does not vanish in this interval. This implies $F(\mu) > 0$ for all $\mu \in (0, \min \{ \mu_*, \nu_* \})$ and thus $\mu_{\text{opt}} = \min \{ \mu_*, \nu_* \}$. There exists a minimal zero μ_0 of $P_{\text{cou}}(\mu)$ in $(0, \min \{ \mu_*, \nu_* \})$: Then $F(\mu)$ is positive definite for $\mu \in (0, \mu_0)$. Hence $F(\mu_0)$ is positive semidefinite and singular. By [30, Prop. 6], $F(\mu)$ cannot be positive definite for $\mu > \mu_0$ and we infer $\mu_{\text{opt}} = \mu_0$. \square

The reader should note the additional simplifications if μ_{pos} and/or ν_{pos} is/are infinite since the functions $P(\cdot)$ and/or $Q(\cdot)$ is/are constant in this case. There is no need to discuss all the details. We just give an interpretation of these results for the model-matching problem introduced in § 2. If one of the matrices $T_2(s)$ or $T_3(s)^T$ has full row rank over $\mathbb{R}(s)$ (the two block problem), either μ_{max} or ν_{max} is infinite. If both matrices have full row rank (the one block problem), the computation of μ_{opt} amounts to the solution of an eigenvalue problem. Theorem 16, however, shows that this may even happen if none of these conditions holds true, i.e., for certain four block problems.

8. Almost disturbance decoupling with stability. Any geometric characterization of $\mu_{\text{opt}} = \infty$ solves the *almost disturbance decoupling problem with \mathbb{C}^- -stability by output measurement* (ADDP). Up until now the ADDP has been treated for *closed* stability sets [37] or the given conditions for the multiple-input multiple-output (MIMO) case are deduced in a rather ad hoc way from single-input single-output (SISO) results without geometric interpretations [21]. Our approach allows us to derive the following new geometric solution of the ADDP in a straightforward manner.

THEOREM 17. μ_{opt} is infinite if and only if

$$(28) \quad \text{im}(G) \subset \mathcal{S}_+(S(s)) \cap \bigcap_{\lambda \in \mathbb{C}^0} \mathcal{S}_\lambda(S(s)),$$

$$(29) \quad \mathcal{V}^+(T(s)) + \sum_{\lambda \in \mathbb{C}^0} \mathcal{V}^\lambda(T(s)) \subset \ker(H),$$

$$(30) \quad \mathcal{V}^+(T(s)) \subset \mathcal{S}_+(S(s)).$$

Proof. If we assume $\mu_{\text{opt}} = \infty$, we obtain (28) from $\mu_* = \infty$, i.e., $\mu_{\text{pos}} = \infty$ and $\mu_{\text{neg}} = \infty$ and Theorem 12. The dual inclusion (29) follows from $\nu_* = \infty$. By $\mu_{\text{pos}} = \infty$ and $\nu_{\text{pos}} = \infty$, the functions $P(\cdot)$ and $Q(\cdot)$ are constant. Therefore $\rho(P(\mu)Q(\mu)) < 1/\mu$ holds for all $\mu \in \mathbb{R}$ if and only if $P(\mu)Q(\mu) = 0$ or $\text{im}(Q(\mu)) \subset \ker(P(\mu))$ are valid for one/all $\mu \in \mathbb{R}$. Theorem 13 together with [31, Thm. 9] implies that the kernel of $P(\mu)$ is given by $\mathcal{S}_+(S(s))$ and, dually, the image of $Q(\mu)$ equals $\mathcal{V}^+(T(s))$. This yields (30). The arguments can be reversed to obtain $\mu_{\text{opt}} = \infty$ from (28), (29), and (30). \square

We should compare this result with the solution of the ADDP for the closed stability set $\mathbb{C}^- \cup \mathbb{C}^0$. The first two conditions differ due to \mathbb{C}^0 -zeros of $S(s)$ and $T(s)$ but the third condition is the same reflecting the earlier statement that the \mathbb{C}^0 -zeros do not cause additional coupling constraints.

The solution of the state-feedback ADDP ($C = I$ and $D = 0$) is contained in the above result: (29) and (30) are automatically satisfied. As it was expected in [21], the relevant space on the right in (28) lies between $\mathcal{S}_+(S(s)) \cap \mathcal{S}_0(S(s)) = \mathcal{V}^-(S(s)) + \mathcal{S}_*(S(s))$ and $\mathcal{S}_+(S(s)) = \mathcal{V}^-(S(s)) + \mathcal{V}^0(S(s)) + \mathcal{S}_*(S(s))$. We stress a simple consequence of our results that cannot be obtained from those in [21]: If $\mu_* = \infty$, it is possible to construct a sequence of *static* stabilizing feedback matrices to approach μ_* .

We finally mention an application of the solution to the ADDP in robust stabilization: As discussed in [16], it is possible to test algebraically whether complete quadratic stabilization is realizable.

Appendix.

Proof of Proposition 15. Since we can decompose any $R_0 \in \mathbb{S}^n$ as $JJ^T - GG^T$, it suffices to prove the result for $R_0 = -GG^T$. Thus we have to show for an arbitrary real $G \in \mathbb{R}^{n \times \times}$ and any $X_0 \in \mathbb{S}^n$:

$$\exists \rho > 0 \quad \exists X > X_0: AX + XA^T + XH^THX - \rho BB^T + GG^T < 0.$$

By Theorem 4, there exist $P > 0$ and F with $(A + BF)^T P + P(A + BF) + PGG^T P + H^T H < 0$. Since

$$(H_\varepsilon \ E_\varepsilon) = \begin{pmatrix} H & 0 \\ 0 & \varepsilon I \end{pmatrix}$$

defines an admissible perturbation of (HE) , we infer from Theorem 8 the existence of some $X > 0$ and ρ_0 with $AX + XA^T + XH^T HX - \rho_0 BB^T + GG^T < 0$. From now on, ρ is always taken out of $[\rho_0, \infty)$ if not specified otherwise. The set

$$\mathcal{X}(\rho) := \{X > 0 \mid AX + XA^T + XH^T HX - \rho BB^T + GG^T < 0\}$$

is nonempty. According to [31, Thm. 9], $\mathcal{X}(\rho)^{-1}$ has a strict lower limit point, which is denoted as $P_\rho \geq 0$ and is, by $\mathcal{X}(\rho) \subset \mathcal{X}(\sigma)$ for $\rho < \sigma$, nonincreasing. Therefore the limit $\lim_{\rho \rightarrow \infty} P_\rho$ exists. If we show that this limit vanishes, we have proved the proposition. Reference [31, Thm. 9] also contains an explicit formula for P_ρ . We transform with some nonsingular S as

$$SAS^{-1} = \begin{pmatrix} A_1 & 0 \\ K_2 H_1 & A_2 \end{pmatrix}, \quad SB = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}, \quad SG = \begin{pmatrix} G_1 \\ G_2 \end{pmatrix}, \quad HS^{-1} = (H_1 \ 0),$$

such that

$$\begin{pmatrix} A_1 - sI \\ H_1 \end{pmatrix}$$

only has unobservable modes in \mathbb{C}^+ and $\sigma(A_2) \subset \mathbb{C}^0 \cup \mathbb{C}^-$. Then P_ρ is equal to

$$S^T \begin{pmatrix} X_\rho^{-1} & 0 \\ 0 & 0 \end{pmatrix} S,$$

where $X_\rho > 0$ is the unique matrix which satisfies

$$A_1 X_\rho + X_\rho A_1^T + X_\rho H_1^T H_1 X_\rho + G_1 G_1^T - \rho B_1 B_1^T = 0, \quad \sigma(A_1 + X_\rho H_1^T H_1) \subset \mathbb{C}^+.$$

Note that X_ρ is nondecreasing.

Let us define $\Delta_\rho := X_\rho - X_{\rho_0} \geq 0$ and $\tilde{A}_1 := A_1 + X_{\rho_0} H_1^T H_1$. Then it is simple to verify [32] that Δ_ρ satisfies

$$\tilde{A}_1 \Delta_\rho + \Delta_\rho \tilde{A}_1^T + \Delta_\rho H_1^T H_1 \Delta_\rho - (\rho - \rho_0) B_1 B_1^T = 0, \quad \sigma(\tilde{A}_1^T + \Delta_\rho H_1^T H_1) \subset \mathbb{C}^+.$$

Let us now assume $\rho > \rho_0$. Obviously,

$$\begin{pmatrix} \tilde{A}_1 - \lambda I & B_1 \\ H_1 & 0 \end{pmatrix}$$

still has maximal row rank for all $\lambda \in \mathbb{C}$, which implies that $(\tilde{A}_1 - sI \ B_1)$ is controllable. Hence Δ_ρ has no kernel and is in fact positive definite. Moreover, we obtain

$$\tilde{A}_1^T \Delta_\rho^{-1} + \Delta_\rho^{-1} \tilde{A}_1 + H_1^T H_1 - (\rho - \rho_0) \Delta_\rho^{-1} B_1 B_1^T \Delta_\rho^{-1} = 0, \quad \sigma(\tilde{A}_1 - (\rho - \rho_0) B_1 B_1^T \Delta_\rho^{-1}) \subset \mathbb{C}^-.$$

By LQ-theory we deduce for any $x_0 \in \mathbb{R}^{n_1}$ (with n_1 as the dimension of \tilde{A}_1)

$$x_0^T \Delta_\rho^{-1} x_0 = \inf \int_0^\infty \frac{1}{\rho - \rho_0} u(t)^T u(t) + x(t)^T H_1^T H_1 x(t) dt,$$

where we vary F such that $\tilde{A}_1 + B_1 F$ is stable and define $x(\cdot)$, $u(\cdot)$ via $\dot{x}(t) = (\tilde{A}_1 + B_1 F)x(t)$, $x(0) = x_0$, $u(t) = Fx(t)$. Again by Theorem 4, there exists a sequence F_j with $\sigma(\tilde{A}_1 + B_1 F_j) \subset \mathbb{C}^-$ and

$$\int_0^\infty \|H_1 e^{(\tilde{A}_1 + B_1 F_j)t}\|^2 dt \rightarrow 0$$

for $j \rightarrow \infty$. This implies $x_0^T \Delta_\rho^{-1} x_0 \rightarrow 0$ for $\rho \rightarrow \infty$. Since x_0 was arbitrary, we obtain $\Delta_\rho^{-1} \rightarrow 0$ and therefore also $X_\rho^{-1} \rightarrow 0$, i.e., $P_\rho \rightarrow 0$ for $\rho \rightarrow \infty$. \square

Acknowledgments. I would like to thank Siep Weiland from the University of Groningen for reading the first draft of the manuscript and for his never-ending interest in this topic. My thanks also go to all the other members of the Systems and Control Group in Groningen (and especially to Professor J. C. Willems) for their kind hospitality.

REFERENCES

- [1] H. ALING and J. M. SCHUMACHER, *A nine-fold canonical decomposition for linear systems*, Internat. J. Control, 39 (1984), pp. 779–805.
- [2] J. A. BALL and N. COHEN, *Sensitivity minimization in an H_∞ norm: Parametrization of all suboptimal solutions*, Internat. J. Control, 46 (1987), pp. 785–816.
- [3] 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.
- [4] J. DOYLE and K. GLOVER, *State-space formulae for all stabilizing controllers that satisfy an H_∞ norm bound and relations to risk sensitivity*, Systems Control Lett., 11 (1988), pp. 167–172.
- [5] J. DOYLE, K. GLOVER, P. KHARGONEKAR, and B. FRANCIS, *State-space solutions to standard H_∞ and H_2 control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831–847.
- [6] L. E. FAIBUSOVICH, *Algebraic Riccati equation and symplectic algebra*, Internat. J. Control, 43 (1986), pp. 781–792.
- [7] ———, *Matrix Riccati inequality: existence of solutions*, Systems Control Lett., 9 (1987), pp. 59–64.
- [8] B. A. FRANCIS, *A Course in H_∞ Control Theory*, Lecture Notes in Computer and Information Science, 88, Springer-Verlag, Berlin, New York, 1987.
- [9] B. A. FRANCIS and J. C. DOYLE, *Linear control theory with an H_∞ optimality criterion*, SIAM J. Control Optim., 25 (1987), pp. 815–844.
- [10] I. GOHBERG, P. LANCASTER, and L. RODMAN, *On Hermitian solutions of the symmetric algebraic Riccati equation*, SIAM J. Control Optim., 24 (1986), pp. 1323–1334.
- [11] M. GREEN, K. GLOVER, D. LIMEBEER, and J. DOYLE, *A J -spectral factorization approach to H_∞ control*, SIAM J. Control Optim., 28 (1990), pp. 1350–1371.
- [12] S. HARA, T. SUGIE, and R. KONDO, *Descriptor form solution for H_∞ control problem with $j\omega$ -axis zeros*, submitted.
- [13] D. HINRICHSSEN and A. J. PRITCHARD, *A robustness measure for linear systems under structured real parameter perturbations*, Report No. 184, Universität Bremen, 1989.
- [14] E. A. JONCKHEERE, J. C. JUANG, and L. M. SILVERMAN, *Spectral theory of the linear-quadratic and H^∞ problems*, Linear Algebra Appl., 122–124 (1989), pp. 273–300.
- [15] P. P. KHARGONEKAR, I. R. PETERSEN, and M. A. ROTEA, *H_∞ -optimal control with 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 and H_∞ -optimal control*, University of Minnesota, 1987.
- [17] H. KIMURA, *Conjugation, interpolation and model-matching in H^∞* , Internat. J. Control, 49 (1989), pp. 269–307.
- [18] H. KWAKERNAAK, *A polynomial approach to minimax frequency domain optimization of multivariable feedback systems*, Internat. J. Control, 41 (1986), pp. 117–156.
- [19] H. W. KNOBLOCH and H. KWAKERNAAK, *Lineare Kontrolltheorie*, Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1985.
- [20] D. J. N. LIMEBEER, E. M. KASENALLY, E. JAÏMOUKA, and M. G. SAFONOV, *A characterization of all solutions to the four block general distance problem*, Proc. IEEE CDC, Austin, 1987.
- [21] A. LINNEMANN, I. POSTLETHWAITE, and B. D. O. ANDERSON, *Almost disturbance decoupling with stabilization by measurement feedback*, Systems Control Lett., 12 (1989), pp. 225–234.

- [22] A. S. MORSE, *Structural invariants of linear multivariable systems*, SIAM J. Control, 11 (1973), pp. 446–465.
- [23] K. M. NAGPAL AND P. P. KHARGONEKAR, *Filtering and smoothing in an H^∞ setting*, Dept. Elect. Engrg., University of Michigan, 1989, preprint.
- [24] I. R. PETERSEN, *Disturbance attenuation and H_∞ optimization: a design method based on the algebraic Riccati equation*, IEEE Trans. Automat. Control, 32 (1987), pp. 427–429.
- [25] A. C. M. RAN AND R. VREUGDENHIL, *Existence and comparison theorems for algebraic Riccati equations for continuous- and discrete-time systems*, Linear Algebra Appl., 99 (1988), pp. 63–83.
- [26] 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.
- [27] M. SAMPEI, T. MITA, AND M. NAKAMICHI, *An algebraic approach to H_∞ output feedback control problems*, Systems Control Lett., 14 (1990), pp. 13–24.
- [28] C. SCHERER, *Almost disturbance decoupling with stability by dynamic output feedback: A sufficient condition*, Systems Control Lett., 10 (1988), pp. 291–299.
- [29] ———, *H_∞ -control by state-feedback: An iterative algorithm and characterization of high-gain occurrence*, System Control Lett., 12 (1989), pp. 383–391.
- [30] ———, *H_∞ -control by state-feedback and fast algorithms for the computation of optimal H_∞ -norms*, IEEE Trans. Automat. Control, 35 (1990), pp. 1090–1099.
- [31] ———, *H_∞ -control by state-feedback for plants with zeros on the imaginary axis*, SIAM J. Control Optim., 30(1992), this issue, pp. 123–142.
- [32] ———, *The solution set of the algebraic Riccati equation and the algebraic Riccati inequality*, Linear Algebra Appl., 153 (1991), pp. 99–122.
- [33] A. A. STOOORVOGEL, *The singular H_∞ control problem with dynamic measurement feedback*, SIAM J. Control Optim., 29 (1991), pp. 160–184.
- [34] ———, *The H_∞ control problem: A state space approach*, Ph.D. thesis, Eindhoven University of Technology, the Netherlands, 1990.
- [35] G. TADMOR, *H_∞ in the time domain: the standard four blocks problem*, Math. of Controls, Signals and Systems, to appear.
- [36] H. L. TRENTELMAN, *Almost invariant subspaces and high gain feedback*, CWI Tracts, Vol. 29, Amsterdam, 1986.
- [37] S. WEILAND AND J. C. WILLEMS, *Almost disturbance decoupling with internal stability*, IEEE Trans. Automat. Control, 34 (1989), pp. 277–286.
- [38] J. C. WILLEMS, *Least-squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, 21 (1971), pp. 319–338.
- [39] ———, *Almost invariant subspaces: An approach to high gain feedback design Part I: Almost controlled invariant subspaces*, IEEE Trans. Automat. Control, 26 (1981), pp. 235–252.
- [40] H. K. WIMMER, *Monotonicity of maximal solutions of algebraic Riccati equations*, Systems Control Lett., 5 (1985), pp. 317–319.
- [41] G. ZAMES, *Feedback and optimal sensitivity: Model reference transformations, multiplicative seminorms and approximate inverses*, IEEE Trans. Automat. Control, 26 (1981), pp. 301–320.
- [42] K. ZHOU AND P. P. KHARGONEKAR, *An algebraic Riccati equation approach to H_∞ optimization*, System Control Lett., 11 (1988), pp. 85–91.

BOUNDARY VELOCITY CONTROL OF INCOMPRESSIBLE FLOW WITH AN APPLICATION TO VISCOUS DRAG REDUCTION*

MAX D. GUNZBURGER†, LISHENG HOU‡, AND THOMAS P. SVOBODNY§

Abstract. An optimal boundary control problem for the Navier–Stokes equations is presented. The control is the velocity on the boundary, which is constrained to lie in a closed, convex subset of $H^{1/2}$ of the boundary. A necessary condition for optimality is derived. Computations are done when the control set is actually finite-dimensional, resulting in an application to viscous drag reduction.

Key words. optimal control, Navier–Stokes equations, boundary control, finite element methods, distributed parameter systems

AMS(MOS) subject classifications. 49A22, 49B22, 49D05, 65N30, 76D05, 93C10, 93C20

1. Introduction. We are concerned with a constrained optimization problem for steady fluid flow, namely that of computing a boundary value of the velocity that minimizes a volume integral that represents frictional energy dissipation. The constraint is the system of equations for viscous incompressible flow. The boundary value of the velocity, hereupon dubbed the control, is constrained to a closed, convex subset of $\mathbf{H}^{1/2}(\Gamma_c)$, where Γ_c is a portion of the body boundary. Such a constraint is necessary since the control does not appear explicitly in the cost functional. Moreover, we cannot eliminate the control in solving coupled state-adjoint equations. In the present case a minimum principle gives us a variational inequality that couples the system of state and adjoint equations. If we further constrain the control set to be in a finite-dimensional subspace, then we are led to an optimization problem where the feasibility set is the set of vertices of a cube in m -dimensional Euclidean space.

Our choice of the cost functional to be optimized is motivated by the following physical consideration (cf. [9]). If a body with boundary Γ is immersed in a fluid, then the force acting on the body is

$$\mathbf{F} = \int_{\Gamma} \mathbf{T} \cdot \mathbf{n} \, d\Gamma,$$

where \mathbf{T} is the stress tensor and \mathbf{n} is the unit normal vector pointing into the body. If we consider a body moving with rectilinear velocity \mathbf{V} , then a measure of the component of the force in the direction of the velocity is

$$\mathbf{F}_D = \mathbf{F} \cdot \mathbf{V} = \mathbf{V} \cdot \int_{\Gamma} \mathbf{T} \cdot \mathbf{n} \, d\Gamma = \int_{\Gamma} \mathbf{v} \cdot \mathbf{T} \cdot \mathbf{n} \, d\Gamma,$$

*Received by the editors March 13, 1989; accepted for publication (in revised form) October 26, 1990.

†Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061. The work of this author was supported by Air Force Office of Scientific Research grants AFOSR-88-0197 and AFOSR-90-0179.

‡Département de Mathématiques et de Statistique, Université Laval, Québec, G1K 7P4. The work of this author was supported by the Department of Education of the Province of Québec, Actions Structurantes Program.

§Department of Mathematics and Statistics, Wright State University, Dayton, Ohio 45435. The work of this author was supported by Air Force Office of Scientific Research grants AFOSR-86-0085 and AFOSR-85-0263 while he was visiting Virginia Tech.

where $\mathbf{v} = \mathbf{v}(\mathbf{x}, t)$ is the velocity field of the fluid. If we assume the linear constitutive law, i.e., an expression for stress is given by

$$\mathbf{T} = -p\mathbf{I} + 2\mu\mathbf{D},$$

where p denotes a pressure field and μ the (constant) viscosity of the fluid, and $\mathbf{D} = \mathbf{D}(\mathbf{v}) = \frac{1}{2}(\nabla\mathbf{v} + \nabla\mathbf{v}^T)$ is a symmetric tensor of first-order derivatives of \mathbf{v} , an integration by parts gives

$$\mathbf{F}_D = 2\mu \int_{\Omega} \mathbf{D} : \mathbf{D} \, dx - \int_{\Omega} p(\operatorname{div} \mathbf{v}) \, dx + \mu \int_{\Omega} \mathbf{v} \cdot \nabla(\operatorname{div} \mathbf{v}) \, dx.$$

(We are assuming that the fluid is isotropic, reflected in the fact that the medium itself is described by the scalar quantity μ .) Although this expression for the drag was derived in the context of the ideal conditions mentioned, it seems reasonable to use the right-hand side as a cost functional to be minimized subject to the constraints imposed by the incompressible Navier–Stokes equations. Of course, on this constraint set, the second and third terms on the right-hand side are clearly zero, thus making our cost functional positive. The term to which this functional then reduces is known in the literature as the *dissipation function*; it represents the rate at which heat energy is conducted into the fluid, or equivalently, the rate at which heat is generated by deformations of the velocity field.

We apply our results to the following problem. It is known that for an aerodynamic body moving at uniform velocity, the main contribution to retardation is the frictional force. This force is increased if the boundary layer becomes turbulent. In addition, the body may have to overcome adverse pressure gradients if the boundary layer separates. We want to reduce viscous (skin-friction) drag of an immersed body whose relative velocity with respect to the fluid is fixed (\mathbf{v}_{∞}). In the specific application that is considered at the end, the method of control is the following: we have m disjoint regions (holes) on the surface of the body where we can specify positive (blowing) or negative (sucking) velocities [4]. Of course, the rate of flow at each hole is strictly limited, not only for the obvious reasons, but because too strong a control would change the lift, and the problem would not represent the one posed. On the other hand, the control velocity must dominate pressure gradients. We will assume that the control set is small enough so that the velocity profile at each hole is fixed, and thus the problem reduces to a finite-dimensional control problem.

The numerical approximation is carried out using the finite-element method; a sketch of the procedure as well as a model problem is given at the end. Further details concerning the approximation of such optimization problems as well as the more difficult (from the viewpoint of approximation) unconstrained case, including error estimates, are given in [3]. The authors are presently applying the techniques and algorithms discussed here to other applications, e.g., flows about airfoils.

1.1. Notation. Real k -dimensional Euclidean space is denoted R^k ; R_+^k is the subset with nonnegative coordinates. The domain Ω is a bounded smooth domain in R^2 or R^3 , whose boundary consists of two connected components $\Gamma = \Gamma_e \cup \Gamma_c$. Furthermore, $\Gamma_e = \Gamma_{(1)} \cup \Gamma_{(2)}$, where $\Gamma_{(1)}$ must have an interior, but $\Gamma_{(2)}$ may be empty. Let $\mathbf{H}^m(B)$ be the Hilbert space of R^n -valued functions defined on B whose j th derivatives, $0 \leq j \leq m$ are in $L^2(B)$. The norm in this space will be denoted by $\|\cdot\|_m$. In all cases, boldface indicates vector-valued. We will use the quotient spaces $L_0^2(\Omega) = \{f \in L^2(\Omega) : \int_{\Omega} f \, dx = 0\}$ and $\mathbf{H}_0^{1/2}(G) = \{\mathbf{g} \in \mathbf{H}^{1/2}(G) : \int_G \mathbf{g} \cdot \mathbf{n} \, dG = 0\}$.

$\langle \cdot, \cdot \rangle_X$ will mean the duality pairing on $X' \times X$, while $\langle\langle \cdot, \cdot \rangle\rangle_X$ will mean the inner product in the Hilbert space X .

The matrix $\{\partial u_i / \partial x_j\}$ will be denoted $\nabla \mathbf{u}$. The velocity deformation tensor is $\mathbf{D}(\mathbf{u}) = \frac{1}{2}(\nabla \mathbf{u} + \nabla \mathbf{u}^T)$. We use the forms

$$\begin{aligned} a(\mathbf{u}, \mathbf{v}) &= 2\nu \int_{\Omega} \mathbf{D}(\mathbf{u}) : \mathbf{D}(\mathbf{v}) \, dx \quad \forall \mathbf{u}, \mathbf{v} \in \mathbf{H}^1(\Omega), \\ b(\mathbf{u}, p) &= - \int_{\Omega} p(\operatorname{div} \mathbf{u}) \, dx \quad \forall p \in L_0^2(\Omega), \quad \mathbf{u} \in \mathbf{H}^1(\Omega), \\ c(\mathbf{u}, \mathbf{v}, \mathbf{w}) &= \int_{\Omega} \mathbf{u} \cdot \nabla \mathbf{v} \cdot \mathbf{w} \, dx \quad \forall \mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbf{H}^1(\Omega). \end{aligned}$$

The following estimate, which is a straightforward application of the Hölder and Sobolev inequalities (cf. [11]), will prove to be useful:

$$|c(\mathbf{u}, \mathbf{v}, \mathbf{w})| \leq \begin{cases} \text{const.} \|\mathbf{u}\|_{1/2} \|\mathbf{v}\|_1 \|\mathbf{w}\|_1 \\ \text{const.} \|\mathbf{u}\|_1 \|\mathbf{v}\|_1 \|\mathbf{w}\|_{1/2} \end{cases}.$$

These forms induce the following operators.

$$A : \mathbf{H}^1(\Omega) \rightarrow \mathbf{H}^{-1}(\Omega)$$

defined by

$$\begin{aligned} \langle A\mathbf{u}, \mathbf{v} \rangle &= a(\mathbf{u}, \mathbf{v}), \quad \mathbf{u} \in \mathbf{H}^1(\Omega), \quad \mathbf{v} \in \mathbf{H}_0^1(\Omega), \\ B : \mathbf{H}^1(\Omega) &\rightarrow L_0^2(\Omega) \end{aligned}$$

defined by

$$\begin{aligned} \langle B\mathbf{u}, p \rangle &= b(\mathbf{u}, p), \quad \mathbf{u} \in \mathbf{H}^1(\Omega), \quad p \in L_0^2(\Omega), \\ \overline{B} : L_0^2(\Omega) &\rightarrow \mathbf{H}^{-1}(\Omega) \end{aligned}$$

defined by

$$\langle \overline{B}p, \mathbf{u} \rangle = b(\mathbf{u}, p) \quad \mathbf{u} \in \mathbf{H}_0^1(\Omega), \quad p \in L_0^2(\Omega),$$

and

$$C : \mathbf{H}^1(\Omega) \times \mathbf{H}^1(\Omega) \rightarrow \mathbf{H}^{-1}(\Omega)$$

defined by

$$\langle C(\mathbf{u}, \mathbf{v}), \mathbf{w} \rangle = c(\mathbf{u}, \mathbf{v}, \mathbf{w}), \quad \mathbf{u}, \mathbf{v} \in \mathbf{H}^1(\Omega), \quad \mathbf{w} \in \mathbf{H}_0^1(\Omega).$$

For theorems concerning operators associated with the stationary Navier–Stokes equations, consult [7]. We will use the trace operators

$$\gamma_e : \mathbf{H}^1(\Omega) \rightarrow \mathbf{H}^{1/2}(\Gamma_e), \quad \gamma_c : \mathbf{H}^1(\Omega) \rightarrow \mathbf{H}^{1/2}(\Gamma_c)$$

and $\gamma = \gamma_e \times \gamma_c$. Optimal solutions of the optimization problem will be tagged with asterisks, \mathbf{v}^* , or with hats, $\hat{\mathbf{v}}$.

2. The optimization problem: boundary velocity control with constraints. We recall that we are faced with the problem of minimizing the functional (\mathbf{F}_D) , subject to the constraint

$$-\nu \Delta \mathbf{v} + (\mathbf{v} \cdot \nabla) \mathbf{v} + \nabla p = 0$$

$$\nabla \cdot \mathbf{v} = 0$$

and either of the sets of boundary conditions

$$(BVC1) \quad \mathbf{v}|_{\Gamma_e} = \mathbf{v}_\infty, \quad \text{and} \quad \mathbf{v}|_{\Gamma_c} = \mathbf{g},$$

(the Boundary Value Control Problem 1), or

$$(BVC2) \quad \mathbf{v}|_{(\Gamma_1)} = \mathbf{v}_\infty, \quad \mathbf{T}(\mathbf{v}) \cdot \mathbf{n}|_{(\Gamma_2)} = 0, \quad \text{and} \quad \mathbf{v}|_{\Gamma_c} = \mathbf{g},$$

(the Boundary Value Control Problem 2), where the control function \mathbf{g} is to lie in \mathcal{U} , a closed, convex subset of $\mathbf{H}^{1/2}(\Gamma_c)$, and where ν is the kinematic viscosity.

Our program is a straightforward approach to a constrained optimization problem: we investigate existence of optimal solutions, we verify the existence of Lagrange multipliers, and then we derive a minimum principle that couples the Euler–Lagrange equations as necessary conditions for the optimal boundary velocity. The cost functional is given by

$$J(\mathbf{v}) = \frac{\nu}{2} \int_{\Omega} |\nabla \mathbf{v} + \nabla \mathbf{v}^T|^2 dx.$$

2.1. Existence of an optimal solution. Consider the Navier–Stokes equations in the weak form,

$$(2.1) \quad a(\mathbf{v}, \mathbf{w}) + c(\mathbf{v}, \mathbf{v}, \mathbf{w}) + b(\mathbf{w}, p) = 0 \quad \forall \mathbf{w} \in \mathbf{H}_0^1(\Omega)$$

$$(2.2) \quad b(\mathbf{v}, q) = 0 \quad \forall q \in L_0^2(\Omega)$$

$$(2.3) \quad \mathbf{v}|_{\Gamma_c} = \mathbf{g}$$

and either

$$(2.4) \quad \mathbf{v}|_{\Gamma_e} = \mathbf{v}_\infty$$

or

$$(2.5) \quad \mathbf{v}|_{(\Gamma_1)} = \mathbf{v}_\infty \quad \mathbf{T}(\mathbf{v}) \cdot \mathbf{n}|_{(\Gamma_2)} = 0.$$

Always, $\mathbf{v}_\infty \in \mathbf{H}^{1/2}(\Gamma_e)$ (or $\mathbf{v}_\infty \in \mathbf{H}^{1/2}(\Gamma_{(1)})$ in the case of (BVC2)) and, in the case of (BVC1), we assume

$$\int_{\Gamma_c} \mathbf{g} \cdot \mathbf{n} d\Gamma + \int_{\Gamma_e} \mathbf{v}_\infty \cdot \mathbf{n} d\Gamma = 0.$$

We consider the case (BVC1), and for simplicity assume that

$$\int_{\Gamma_e} \mathbf{v}_\infty \cdot \mathbf{n} d\Gamma = 0,$$

so that

$$\int_{\Gamma_c} \mathbf{g} \cdot \mathbf{n} d\Gamma = 0.$$

(Cases (BVC2) and (BVC1) without the above assumption can be treated in similar manner.)

DEFINITION 1. We define the *admissibility set*

$$\mathcal{T}_{\text{ad}} = \{(\mathbf{v}, \mathbf{g}) \in \mathbf{H}^1(\Omega) \times \mathcal{U}(\subseteq \mathbf{H}^{1/2}(\Gamma_c)) : J(\mathbf{v}) < \infty \text{ and there exists } p \in L_0^2(\Omega) \text{ and so that (2.2)–(2.4) and either (2.4) or (2.5) are satisfied}\}.$$

DEFINITION 2. We define an *optimal solution* $(\mathbf{v}^*, \mathbf{g}^*)$ to be one for which $J(\mathbf{v}^*) \leq J(\mathbf{v})$ for any $(\mathbf{v}, \mathbf{g}) \in \mathcal{T}_{\text{ad}}$.

DEFINITION 3. We define a *local minimum* $(\hat{\mathbf{v}}, \hat{\mathbf{g}})$ to be one for which there exists $\epsilon > 0$ such that $J(\hat{\mathbf{v}}) \leq J(\mathbf{v})$ for any $(\mathbf{v}, \mathbf{g}) \in \mathcal{T}_{\text{ad}}$ with $\|\mathbf{v} - \hat{\mathbf{v}}\|_1 < \epsilon$

LEMMA 2.1. *There exists an optimal solution to the above problem.*

Proof. First, note that \mathcal{T}_{ad} is nonempty. This follows from the fact

$$J(\mathbf{v}) \leq \nu \|\mathbf{v}\|_1^2$$

and well-known existence theorems for solutions of stationary Navier–Stokes equations (cf. [1], [8], and [10]). Now let $\mathbf{V}_m = (\mathbf{v}_m, \mathbf{g}_m)$ be a sequence in \mathcal{T}_{ad} such that

$$\lim_{m \rightarrow \infty} J(\mathbf{v}_m) = \inf_{v \in \mathcal{T}_{\text{ad}}} J(\mathbf{v}).$$

By definition of \mathcal{T}_{ad} , $J(\mathbf{v}_m) \leq C_0$, a constant independent of m . On the boundary, $\mathbf{v} = \mathbf{v}_\infty \in \mathbf{H}^{1/2}(\Gamma_e)$, and because of the assumption on Γ_e we have

$$J(\mathbf{v}) \geq c_1 \|\mathbf{v}\|_1^2 - c_2 \|\mathbf{v}\|_1 - c_3$$

(this is a direct consequence of Korn's inequality, cf. [2, p.117]); thus, $\|\mathbf{v}_m\|_1 \leq C_V$. By the trace theorem, $\|\mathbf{g}_m\|_{1/2} \leq C_U$. We therefore have the weak limits

$$\mathbf{v}_m \rightharpoonup \mathbf{v}^* \in \mathbf{H}^1(\Omega), \quad \mathbf{g}_m \rightharpoonup \mathbf{g}^* \in \mathbf{H}^{1/2}(\Gamma_C).$$

Since $\mathbf{H}^1(\Omega)$ imbeds compactly in $\mathbf{L}^2(\Omega)$, we also have the strong limit

$$\mathbf{v}_m \rightarrow \mathbf{v}^* \in \mathbf{L}^2(\Omega).$$

As \mathcal{U} is convex, closed, it is closed in the weak topology, thus $\mathbf{g}^* \in \mathcal{U}$. Next, using the fact that J is bounded, convex, we have that it is weakly lower semicontinuous, and thus $J(\mathbf{v}^*) = \inf_{v \in \mathcal{T}_{\text{ad}}} J(\mathbf{v})$.

Now we must show that the limit \mathbf{v}^* is in the admissible set, \mathcal{T}_{ad} . We first show that \mathbf{v}^* satisfies the boundary conditions. Since

$$\gamma_e v_m = \mathbf{v}_\infty, \quad \gamma_c \mathbf{v}_m = \mathbf{g}_m,$$

we want to show that $\gamma_e \times \gamma_c$ is a weakly closed operator. But it is continuous on $\mathbf{H}^1(\Omega)$ and so it is closed, and thus has a closed graph, and thus by the Hahn–Banach theorem, is weakly closed. Thus

$$\gamma_e \mathbf{v}^* = \mathbf{v}_\infty, \quad \gamma_c \mathbf{v}^* = \mathbf{g}^*.$$

Now we show that the limiting flow is incompressible. The condition

$$b(\mathbf{v}_m, q) = 0, \quad \forall q \in \mathbf{L}^2$$

means that $\mathbf{v}_m \in \ker(B)$ for all m . But since $\ker(B)$ is a closed subspace, it is weakly closed, and so

$$\mathbf{v}^* \in \ker(B) \Rightarrow b(\mathbf{v}^*, q) = 0 \quad \forall q \in L^2(\Omega).$$

It is clear that $a(\mathbf{v}_m, \mathbf{w}) \rightarrow a(\mathbf{v}^*, \mathbf{w})$ for all $\mathbf{w} \in \mathbf{H}^1(\Omega)$. On the other hand, the operator $C(\cdot, \cdot)$ is weak sequentially continuous. To see this, take $\mathbf{w} \in \mathbf{C}_0^\infty(\Omega)$, and consider

$$c(\mathbf{v}_m, \mathbf{v}_m, \mathbf{w}) = -c(\mathbf{v}_m, \mathbf{w}, \mathbf{v}_m) = - \sum_{i,j} \int_{\Omega} v_{mi} \frac{\partial w_j}{\partial x_i} v_{mj} \, dx.$$

Since $(\nabla \mathbf{w})_{ij} \in L^\infty(\Omega)$, and since $\mathbf{v}_m \rightarrow \mathbf{v}^*$ in $\mathbf{L}^2(\Omega)$, we have

$$c(\mathbf{v}_m, \mathbf{v}_m, \mathbf{w}) \rightarrow c(\mathbf{v}^*, \mathbf{v}^*, \mathbf{w}) \quad \forall \mathbf{w} \in \mathbf{C}_0^\infty;$$

but this space is dense in $\mathbf{H}_0^1(\Omega)$, thus

$$A\mathbf{v}_m + C(\mathbf{v}_m, \mathbf{v}_m) \rightarrow A\mathbf{v}^* + C(\mathbf{v}^*, \mathbf{v}^*) \in \mathbf{H}^{-1}.$$

Moreover, $A\mathbf{v}_m + C(\mathbf{v}_m, \mathbf{v}_m) \in R(\overline{B})$ for all m ; but the inf sup condition ([7, p.81]) implies that $R(\overline{B})$ is closed, and thus

$$A\mathbf{v}^* + C(\mathbf{v}^*, \mathbf{v}^*) \in R(\overline{B}),$$

or, in other words, there exists a $p^* \in L_0^2(\Omega)$ such that

$$A\mathbf{v}^* + C(\mathbf{v}^*, \mathbf{v}^*) = -\overline{B}p^*,$$

which completes the proof of existence of an optimal control. \square

2.2. The Lagrange multiplier rule. The idea of the technique of Lagrange multipliers is that the problem of minimizing a functional subject to a constraint can be reduced to the unconstrained minimization of an auxiliary functional, the Lagrangian. Important in applications of convex optimization is the existence of a minimum principle. (The general theory of the parametrized Lagrange multiplier rule is developed in [5, Chap. 1] and [13, Chap. 1].)

We can define a mapping

$$\mathbf{F} : \mathbf{H}^1(\Omega) \times L_0^2(\Omega) \times \mathcal{U}_{\text{ad}} \rightarrow \mathbf{H}^{-1}(\Omega) \times L_0^2(\Omega) \times \mathbf{H}^{1/2}(\Gamma_e) \times \mathbf{H}^{1/2}(\Gamma_c)$$

by

$$F^1(\mathbf{v}, p, \mathbf{g}) = A\mathbf{v} + C(\mathbf{v}, \mathbf{v}) + \overline{B}p \in \mathbf{H}^{-1}(\Omega)$$

$$F^2(\mathbf{v}, p, \mathbf{g}) = B\mathbf{v} \in L_0^2(\Omega)$$

$$F^3(\mathbf{v}, p, \mathbf{g}) = \gamma_e \mathbf{v} - \mathbf{v}_\infty \in \mathbf{H}^{1/2}(\Gamma_e)$$

$$F^4(\mathbf{v}, p, \mathbf{g}) = \gamma_c \mathbf{v} - \mathbf{g} \in \mathbf{H}^{1/2}(\Gamma_c)$$

For economy, we say $\mathbf{F} : \mathbf{X} \times \mathcal{U} \rightarrow \mathbf{Y}$. Thus, our constraint (2.1)-(2.4) is

$$\mathbf{F} = 0 \in \mathbf{H}^{-1}(\Omega) \times L_0^2(\Omega) \times \mathbf{H}^{1/2}(\Gamma_e) \times \mathbf{H}^{1/2}(\Gamma_c)$$

or

$$\mathbf{F} = 0 \in \mathbf{Y}.$$

We say that the set \mathcal{U} has *property C* at $(\hat{\mathbf{v}}, \hat{\mathbf{g}})$ if for any nonzero solution $(\boldsymbol{\xi}, \sigma)$ of the system

$$(2.6) \quad \begin{aligned} -\nu \delta \xi_i + \xi_j \frac{\partial \hat{v}_j}{\partial x_1} - \hat{v}_j \frac{\partial \xi_i}{\partial x_j} + \frac{\partial \sigma}{\partial x_i} &= 0 \\ \nabla \cdot \boldsymbol{\xi} &= 0 \\ \boldsymbol{\xi}|_{\Gamma} &= 0 \end{aligned}$$

we can find $\mathbf{g} \in \mathcal{U}$ such that

$$(2.7) \quad \int_{\Gamma_C} \mathbf{T}(\boldsymbol{\xi}) \cdot \mathbf{n} \cdot (\mathbf{g} - \hat{\mathbf{g}}) d\Gamma < 0.$$

Convention will have it that property C is to hold vacuously if there are no nonzero solutions of (2.6).

Remark. As in [3] we can show that $\mathcal{U} = \mathbf{H}^{1/2}(\Gamma)$ has property C (this is the case where the control is not constrained). This conclusion is seen to be equivalent to the following uniqueness result: If (2.7) is not true, then $\mathbf{T}(\boldsymbol{\xi}) = 0$ and if $\boldsymbol{\xi}$ is a solution to (2.6), then $\boldsymbol{\xi} = 0$.

We will use the notation $D_1 \mathbf{F}(\hat{\mathbf{v}}, \hat{\mathbf{g}}) = D_{(u,p)} \mathbf{F}(\hat{\mathbf{v}}, \hat{p}, \hat{\mathbf{g}}) = \hat{F}$, $\mathbf{W} = (\mathbf{w}, q)$, $\mathbf{Z} = \text{image}(\hat{F})$, $\mathbf{F}(\mathbf{v}, p, \mathcal{U}) = \text{image of the map } \mathbf{g} \rightarrow \mathbf{F}(\mathbf{v}, p, \mathbf{g}), \mathbf{g} \in \mathcal{U}$, $\mathbf{F}(\mathcal{U}) = \mathbf{F}(\hat{\mathbf{v}}, \hat{p}, \mathcal{U})$; note that $\mathbf{F}(\mathcal{U})$ is convex since \mathbf{F} depends linearly on the control.

We are now in a position to state the main result of this section.

THEOREM 1. *Let $(\hat{\mathbf{v}}, \hat{\mathbf{g}})$ be a local minimum (in the sense of Definition 3). Then there exists a nonzero Lagrange multiplier,*

$$\begin{aligned} \mathbb{Z} = (\boldsymbol{\xi}, \sigma, \boldsymbol{\zeta}_e, \boldsymbol{\zeta}_c) &\in [\mathbf{H}^{-1}(\Omega) \times L_0^2(\Omega) \times \mathbf{H}^{1/2}(\Gamma_e) \times \mathbf{H}^{1/2}(\Gamma_c)]' \\ &= \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega) \times \mathbf{H}^{-1/2}(\Gamma_e) \times \mathbf{H}^{-1/2}(\Gamma_c) \\ &(\text{or } = \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega) \times (\mathbf{H}_0^{1/2}(\Gamma_{(1)}))' \times (\mathbf{H}_0^{1/2}(\Gamma_c))') \end{aligned}$$

(the dual spaces depend on the specified (BVP)), so that if \mathcal{U} satisfies property C at $(\hat{\mathbf{v}}, \hat{\mathbf{g}})$, then $\hat{\mathbb{Z}}$ solves the variational system

$$(2.8) \quad \langle J'(\hat{\mathbf{v}}), \mathbf{w} \rangle + \langle \hat{\mathbb{Z}}, D_1 \mathbf{F}(\hat{\mathbf{v}}, \hat{\mathbf{g}})(\mathbf{w}, q) \rangle = 0, \quad \forall \mathbf{W} \in \mathbf{H}^1(\Omega) \times L_0^2(\Omega).$$

If property C is not valid then $\hat{\mathbb{Z}}$ is a nonzero solution to

$$(2.9) \quad \langle \hat{\mathbb{Z}}, D_1 \mathbf{F}(\hat{\mathbf{v}}, \hat{\mathbf{g}}) \mathbf{W} \rangle = 0, \quad \forall \mathbf{W} \in \mathbf{H}^1(\Omega) \times L_0^2(\Omega);$$

in either case, $\hat{\mathbb{Z}}$ satisfies the variational inequality

$$(2.10) \quad \langle \hat{\mathbb{Z}}, \mathbf{F}(\hat{\mathbf{v}}, \hat{p}, \mathbf{g}) \rangle \geq 0, \quad \forall \mathbf{g} \in \mathcal{U}.$$

The inequality (2.10) will henceforth be referred to as the *minimum principle*. The Lagrangian for this problem is

$$\begin{aligned} L = J(\mathbf{v}) + \langle \mathbb{Z}, \mathbf{F}(\mathbf{v}, p, \mathbf{g}) \rangle &= a(\mathbf{v}, \mathbf{v}) + a(\mathbf{v}, \boldsymbol{\xi}) + c(\mathbf{v}, \mathbf{v}, \boldsymbol{\xi}) \\ &+ b(\boldsymbol{\xi}, p) + b(\mathbf{v}, \sigma) + \langle \boldsymbol{\zeta}_1, \gamma_e \mathbf{v} - \mathbf{v}_\infty \rangle_{\Gamma} + \langle \boldsymbol{\zeta}_2, \gamma_C \mathbf{v} - \mathbf{g} \rangle_{\Gamma}. \end{aligned}$$

The proof of Theorem 1 will rely on the following result.

THEOREM 2. (Ioffe and Tikhomirov [5]). *In the notation of the foregoing, if*

- (a) \mathbf{X} and \mathbf{Y} are Banach spaces;
- (b) $J \in C^1(\mathbf{X})$, $D_1\mathbf{F}(\mathbf{v}, p) \in C^1(\mathbf{X})$;
- (c) $\mathbf{F}(\mathbf{v}, p, \mathcal{U})$ convex;
- (d) \mathbf{Z} is closed in \mathbf{Y} and has finite codimension;
- (e) some condition of complete regularity holds (for instance, $0 \in \mathbf{Y}$ is an interior point of $\mathbf{Z} + \mathbf{F}(\mathcal{U})$);

then there exists a $\hat{\mathbf{Z}}$ as described in Theorem 1 and satisfying (2.8) and (2.10).

Proof. For the proof see [13, Theorem P, p. 49] or [5, Theorem 3, p. 71, p. 85].

□

Proof of Theorem 1. We wish to apply Theorem 2 to construct a \mathbf{Z} that satisfies (2.8) and (2.10). It is clear that conditions (a), (b), and (c) follow from previous observations. To show condition (d) is true, we show that \hat{F} is Fredholm. If we compute the derivative,

$$\hat{F}(\mathbf{w}, q) = D_1 F(\hat{\mathbf{u}}, \hat{p}, \hat{\mathbf{g}}) \begin{pmatrix} \mathbf{w} \\ q \end{pmatrix} = \begin{pmatrix} A\mathbf{w} + \overline{B}q + C(\hat{\mathbf{u}}, \mathbf{w}) + C(\mathbf{w}, \hat{\mathbf{u}}) \\ B\mathbf{w} \\ \gamma_e \mathbf{w} \\ \gamma_c \mathbf{w} \end{pmatrix}.$$

By the trace theorem, and using the ellipticity of A and the inf sup property for B , we can see (cf. [1], [6], and [7]) that the Stokes operator

$$\begin{pmatrix} A & \overline{B} \\ B & 0 \\ \gamma & 0 \end{pmatrix}$$

is an isomorphism from $\mathbf{H}^1(\Omega) \times L_0^2(\Omega) \rightarrow \mathbf{H}^{-1}(\Omega) \times L_0^2(\Omega) \times \mathbf{H}^{1/2}(\Gamma)$. (The last factor in this product of spaces is $\mathbf{H}_0^{1/2}(\Gamma)$ if the velocity is specified on the whole outer boundary, i.e., (BVC1).) We estimate the linear operators $C(\cdot, \hat{\mathbf{u}})$ and $C(\hat{\mathbf{u}}, \cdot)$ as follows. From the estimate

$$|c(\mathbf{w}, \hat{\mathbf{u}}, \boldsymbol{\xi})| \leq \|\mathbf{w}\|_{1/2} \|\hat{\mathbf{u}}\|_1 \|\boldsymbol{\xi}\|_1,$$

we have that $C(\cdot, \hat{\mathbf{u}})$ is continuous from $\mathbf{H}^{1/2}(\Omega)$ into $\mathbf{H}^{-1}(\Omega)$ and thus compact from $\mathbf{H}^1(\Omega)$ into $\mathbf{H}^{-1}(\Omega)$. Likewise, from the estimate

$$|c(\hat{\mathbf{u}}, \mathbf{w}, \boldsymbol{\xi})| \leq \|\hat{\mathbf{u}}\|_1 \|\mathbf{w}\|_1 \|\boldsymbol{\xi}\|_{1/2},$$

we see that $C(\hat{\mathbf{u}}, \cdot)$ is continuous from $\mathbf{H}^1(\Omega)$ into $\mathbf{H}^{-1/2}(\Omega)$ and thus compact from $\mathbf{H}^1(\Omega)$ into $\mathbf{H}^{-1}(\Omega)$. Thus, $\hat{F} = D_1 \mathbf{F}(\hat{\mathbf{u}}, \hat{p}, \hat{\mathbf{g}})$ is a Fredholm operator. Since \mathbf{Z} is now known to be closed and with finite codimension M (being the dimension of the space of solutions of (2.6)), we can write

$$\mathbf{Y} = \mathbf{Z} \oplus \mathbf{E}.$$

Property C expresses a condition of complete regularity (assumption (e) in Theorem 2); to see this, note that, under the assumption that property C holds, if $y \in \mathbf{Y}$, then there exist $(\mathbf{w}_1, q_1) \in \mathbf{X}$, $\mathbf{g} \in \mathcal{U}$, $\lambda \geq 0$ such that

$$(2.11) \quad y = \hat{F}(\mathbf{w}_1, q_1) + \lambda(\hat{\mathbf{v}}, \hat{p}, \mathbf{g}).$$

Let $Q : \mathbf{Y} \rightarrow \mathbf{Y}/\mathbf{Z} \cong \mathbf{E}$ be the canonical mapping. If (2.11) does not hold, then taking into account the finite dimensionality of \mathbf{E} and the convexity of $\mathbf{F}(\mathcal{U})$, we can find $y \in \mathbf{Y}$ such that

$$Qy \cdot Q\mathbf{F}(\hat{\mathbf{v}}, \hat{p}, \mathbf{g}) \leq 0, \quad \forall \mathbf{g} \in \mathcal{U}.$$

Define $\lambda = -Qy/|Qy|$ and $\Lambda = Q^*\lambda$, then clearly, $\Lambda \in \mathbf{Z}^\perp$ and $\langle \Lambda, \mathbf{F}(\mathcal{U}) \rangle \geq 0$, i.e., property C does not hold. Thus, we can find a finite number of $\mathbf{g}_i \in \mathcal{U}$, $1 \leq i \leq N$, such that $\sum \mathbf{F}(\hat{\mathbf{v}}, \hat{p}, \mathbf{g}_i) = \hat{\mathbf{F}}\mathbf{W}'$ for some $\mathbf{W}' = (\mathbf{w}', q') \in \mathbf{X}$, and so that the mapping

$$(\mathbf{w}, q, \lambda_1, \dots, \lambda_n) \mapsto \hat{\mathbf{F}}(\mathbf{w}, q) + \sum \lambda_i \mathbf{F}(\hat{\mathbf{v}}, \hat{p}, \mathbf{g}_i)$$

defined on $\mathbf{X} \times R_+^N$ is onto \mathbf{Y} . Let ϵ be given and let $(\mathbf{w}_0, q_0, \mathbf{g}_0)$ be such that

$$(2.12) \quad \hat{\mathbf{F}}(\mathbf{w}_0, q_0) + \mathbf{F}(\hat{\mathbf{v}}, \hat{p}, \mathbf{g}_0) = 0.$$

Then the mapping

$$\begin{aligned} G_\epsilon(\mathbf{v}, p, \mathbf{a}) = & (1 - a_0 - \epsilon \sum a_i) \mathbf{F}(\hat{\mathbf{v}} + \mathbf{v}, \hat{p} + p, \hat{\mathbf{g}}) \\ & + a_0 \mathbf{F}(\hat{\mathbf{v}} + \mathbf{v}, \hat{p} + p, \mathbf{g}_0) + \epsilon \sum a_i \mathbf{F}(\hat{\mathbf{v}} + \mathbf{v}, \hat{p} + p, \mathbf{g}_i), \end{aligned}$$

defined on $\mathbf{X} \times R_+^{N+1}$, is smooth at the origin and $DG(0, 0, 0)$ is surjective. We may now continue with the proof of Theorem 2 given in [5, p. 87] or [13, p. 51] to construct a $\hat{\mathbf{Z}}$ that satisfies (2.8) and (2.10).

The stationarity condition (2.8) gives

$$\begin{aligned} 0 = & a(\hat{\mathbf{v}}, \mathbf{w}) + a(\mathbf{w}, \hat{\boldsymbol{\xi}}) + c(\hat{\mathbf{v}}, \mathbf{w}, \hat{\boldsymbol{\xi}}) + c(\mathbf{w}, \hat{\mathbf{v}}, \hat{\boldsymbol{\xi}}) + \\ & b(\hat{\boldsymbol{\xi}}, q) + b(\mathbf{w}, \sigma) + \langle \boldsymbol{\zeta}_1, \gamma_e \mathbf{w} \rangle + \langle \boldsymbol{\zeta}_2, \gamma_c \mathbf{w} \rangle. \end{aligned}$$

The minimum principle (2.10), reduces in our case to the variational inequality

$$\langle \boldsymbol{\zeta}_2, \hat{\mathbf{g}} - \mathbf{g} \rangle \geq 0 \quad \forall \mathbf{g} \in \mathcal{U}.$$

To get a (formal) expression for the multiplier $\boldsymbol{\zeta}_2$, we integrate by parts to obtain the system of equations for the adjoint variables

$$\begin{aligned} -\nu \Delta \hat{\xi}_i + \hat{\xi}_j \frac{\partial \hat{v}_j}{\partial x_i} - \hat{v}_j \frac{\partial \hat{\xi}_i}{\partial x_j} + \frac{\partial \hat{\sigma}}{\partial x_i} &= -\nu \Delta \hat{v}_i \\ \nabla \cdot \hat{\boldsymbol{\xi}} &= 0 \\ \hat{\boldsymbol{\xi}}|_\Gamma &= 0 \\ \hat{\boldsymbol{\zeta}}_1 &= -(\mathbf{T}(\hat{\mathbf{v}}) + \mathbf{T}(\hat{\boldsymbol{\xi}}))|_{\Gamma_e} \cdot \mathbf{n} \\ \hat{\boldsymbol{\zeta}}_2 &= -(\mathbf{T}(\hat{\mathbf{v}}) + \mathbf{T}(\hat{\boldsymbol{\xi}}))|_{\Gamma_c} \cdot \mathbf{n}. \end{aligned}$$

Thus, the minimum principle reduces to

$$(2.13) \quad \int_{\Gamma_c} [\mathbf{T}(\hat{\mathbf{v}}) + \mathbf{T}(\hat{\boldsymbol{\xi}})] \cdot \mathbf{n} \cdot (\mathbf{g} - \hat{\mathbf{g}}) d\Gamma \geq 0, \quad \forall \mathbf{g} \in \mathcal{U}.$$

Note that this condition is equivalent to:

choose $\hat{\mathbf{g}} \in \mathcal{U}$ that solves

$$\max_{\hat{\mathbf{g}} \in \mathcal{U}} \min_{\mathbf{g} \in \mathcal{U}} \int_{\Gamma} [\mathbf{T}(\mathbf{v}) + \mathbf{T}(\boldsymbol{\xi})] \cdot \mathbf{n} \cdot (\mathbf{g} - \hat{\mathbf{g}}) d\Gamma.$$

In the case where property C does not hold it follows immediately that there exists

$$\begin{aligned} \mathbb{Z} = (\boldsymbol{\xi}, \boldsymbol{\sigma}, \boldsymbol{\zeta}_e, \boldsymbol{\zeta}_c) &\in [\mathbf{H}^{-1}(\Omega) \times L_0^2(\Omega) \times \mathbf{H}^{1/2}(\Gamma) \times \mathbf{H}^{1/2}(\Gamma)]' \\ &= \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega) \times \mathbf{H}^{-1/2}(\Gamma) \times \mathbf{H}^{-1/2}(\Gamma) \\ (\text{or } &= \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega) \times (\mathbf{H}_0^{1/2}(\Gamma_{(1)}))' \times (\mathbf{H}_0^{1/2}(\Gamma_c))') \end{aligned}$$

that satisfies the system of equations (2.9), i.e.,

$$\begin{aligned} -\nu \Delta \hat{\xi}_i + \hat{\xi}_j \frac{\partial \hat{v}_j}{\partial x_i} - \hat{v}_j \frac{\partial \hat{\xi}_i}{\partial x_j} + \frac{\partial \hat{\sigma}}{\partial x_i} &= 0 \\ \nabla \cdot \hat{\boldsymbol{\xi}} &= 0 \\ \hat{\boldsymbol{\xi}}|_{\Gamma} &= 0 \\ \hat{\boldsymbol{\zeta}}_2 &= -\mathbf{T}(\hat{\boldsymbol{\xi}})|_{\Gamma_c} \cdot \mathbf{n} \end{aligned}$$

and

$$\int_{\Gamma_c} \mathbf{T}(\hat{\boldsymbol{\xi}}) \cdot \mathbf{n} \cdot (\mathbf{g} - \hat{\mathbf{g}}) d\Gamma \geq 0, \quad \forall \mathbf{g} \in \mathcal{U};$$

thus inequality (2.10) is seen to hold in all cases. \square

Let us see what Theorem 1 gives us in the case where \mathcal{U} is given more structure. Suppose that \mathcal{U} is given by

$$\mathcal{U} = \{\mathbf{g} \in \mathbf{H}^{1/2} : \mathbf{g} \in \mathbf{A}(\mathbf{s}), \quad \text{for almost all } \mathbf{s} \in \Gamma_c\}$$

where $\mathbf{A}(\mathbf{s})$ is a compact, convex subset of R^n for all \mathbf{s} . If we consider the sets

$$\mathcal{G}_{s_0, \epsilon} = \{\mathbf{g} \in \mathcal{U} \ni \mathbf{g} = \hat{\mathbf{g}} \text{ on } \Gamma_{s_0, \epsilon} = \Gamma - \Gamma \cap B(\mathbf{s}_0, \epsilon)\},$$

then our condition reduces to

$$\int_{\Gamma_{s_0, \epsilon}} [\mathbf{T}(\hat{\mathbf{v}}) + \mathbf{T}(\hat{\boldsymbol{\xi}})] \cdot \mathbf{n} \cdot (\mathbf{g} - \hat{\mathbf{g}}) d\Gamma \geq 0, \quad \forall \mathbf{g} \in \mathcal{G}_{s_0, \epsilon}, \quad \forall \mathbf{s}_0 \in \Gamma_c, \quad \forall \epsilon > 0.$$

But since $\mathbf{s}_0 \in \Gamma_c, \epsilon > 0$ are arbitrary, we find that the pointwise condition

$$[\mathbf{T}(\hat{\mathbf{v}})(\mathbf{s}) + \mathbf{T}(\hat{\boldsymbol{\xi}})(\mathbf{s})] \cdot \mathbf{n} \cdot (\mathbf{g} - \hat{\mathbf{g}}) \geq 0, \quad \forall \mathbf{g}(\mathbf{s}) \in \mathbf{A}(\mathbf{s}),$$

holds for almost all $\mathbf{s} \in \Gamma_c$.

3. Finite-dimensional control. Suppose that \mathcal{U} is a compact convex subset of T , an m -dimensional subspace of $\mathbf{H}^{1/2}(\Gamma)$ spanned by the m functions $\{\boldsymbol{\phi}^k\}_1^m$. For the sake of unity of presentation we will assume that property C holds, although everything is true mutatis mutandis in any case and the final recipe is the same. Let the set of

vectors $\{\mathbf{w}^k\}$, $1 \leq k \leq m$, be a basis for R^m , which we identify with the set $\{\phi^k\}$, so that T can be thought of as either $R^m = \text{span}\{\mathbf{w}_k\}$ or as $\text{span}\{\phi^k\} \subset \mathbf{H}^{1/2}(\Gamma_c)$. Thus any control vector $\mathbf{g} \in \mathcal{U}$ has a representation of the form

$$\mathbf{g}(\mathbf{s}) = \sum g_k \phi^k \quad \text{or} \quad \mathbf{g} = \sum g_k \mathbf{w}^k$$

in T .

Let $\Pi : \mathbf{H}^{1/2} \rightarrow T$ be the projection. Let $j : T \rightarrow \mathbf{H}^{1/2}$ be the canonical injection. Then we can write the components, $\{\Pi \mathbf{g}\}_k$, as

$$\{\Pi \mathbf{g}\}_k = \langle \langle \mathbf{g}, \phi^k \rangle \rangle_T .$$

Here we use the $\mathbf{H}^{1/2}$ inner product.

Recall the minimum principle (2.13). The integral on the left in that inequality is the duality relation $\mathbf{H}^{-1/2} \times \mathbf{H}^{1/2}$, which can be written

$$\langle (T(\mathbf{v}) + T(\hat{\xi})) \cdot \mathbf{n}, j\Pi(\mathbf{g} - \hat{\mathbf{g}}) \rangle$$

for $\mathbf{g} \in T$. This equals

$$\langle j'(\mathbf{T} + \mathbf{T}), \Pi(\mathbf{g} - \hat{\mathbf{g}}) \rangle$$

which is the duality pairing of $T' \times T$. Thus the minimum principle is

$$\langle j'(\mathbf{T}(\hat{\mathbf{v}}) + \mathbf{T}(\hat{\xi})) \cdot \mathbf{n}, (\mathbf{g} - \hat{\mathbf{g}}) \rangle \geq 0, \quad \forall \mathbf{g} \in \mathcal{U} .$$

(This is duality pairing in $T' \times T$.) Let us give T' the dual basis:

$$\text{either } \{\omega_k\}, \quad \text{or} \quad \{\psi^k\}$$

where

$$\langle \omega_k, \mathbf{w}_j \rangle = \delta_{kj}, \quad \langle \psi^k, \phi^j \rangle = \delta_{kj} .$$

The operator dual to j , $j' : \mathbf{H}^{-1/2} \rightarrow T'$ is a projection since T is closed. Thus we can define components

$$\eta^k = \langle \langle j'(\mathbf{T}(\hat{\mathbf{v}}) + \mathbf{T}(\hat{\xi})), \psi^k \rangle \rangle_{-1/2}$$

so that the minimum principle (2.13) now becomes

$$\sum_k \eta^k (g - \hat{g})_k \geq 0, \quad \forall \mathbf{g} = \{g_k\} \in \mathcal{U} .$$

Let us look at the special case where \mathcal{U} is a parallelepiped

$$\{a_k \leq g_k \leq b_k\} ;$$

then the minimum principle is seen to resolve itself as m independent inequations

$$\eta^k (h - \hat{g}_k) \geq 0, \quad \forall h \in [a_k, b_k]$$

and thus we have the following prescription for the optimal control:

$$(3.1) \quad \begin{aligned} &\text{if } \eta^k > 0 \text{ then } \hat{g}_k = a_k \\ &\text{if } \eta^k < 0 \text{ then } \hat{g}_k = b_k . \end{aligned}$$

If $\eta^k = 0$, then the Lagrange necessary condition gives no information about \hat{g}_k .

We see that in this procedure, the η_k would seem to have to be computed in the $\mathbf{H}^{-1/2}$ inner product, which may pose computational difficulties. It turns out that by choosing the $\{\phi^k\}$ properly we can avoid this situation. In particular, suppose that the $\{\phi^k\}$ form an orthonormal set in $\mathbf{H}^{1/2}$. Let $\Lambda : \mathbf{H}^{1/2} \rightarrow \mathbf{H}^{-1/2}$ be the canonical isomorphism; then

$$\langle \Lambda \phi^k, \phi^j \rangle = \delta_{kj}$$

where this is duality. This means that we can choose the sequence $\{\psi^k\} = \{\Lambda \phi^k\}$, and

$$\begin{aligned} \eta_k &= \langle \mathbf{T}(\mathbf{v}) + \mathbf{T}(\boldsymbol{\xi}), \psi^k \rangle_{-1/2} = \langle \mathbf{T}(\hat{\mathbf{v}}) + \mathbf{T}(\boldsymbol{\xi}), \Lambda \phi^k \rangle_{-1/2} \\ &= \langle \mathbf{T}(\hat{\mathbf{v}}) + \mathbf{T}(\boldsymbol{\xi}), \phi^k \rangle; \end{aligned}$$

in other words,

$$(3.2) \quad \eta^k = \int_{\Gamma_c} [\mathbf{T}(\hat{\mathbf{v}}) + \mathbf{T}(\hat{\boldsymbol{\xi}})] \cdot \mathbf{n} \cdot \phi^k d\Gamma.$$

In the degenerate case, the prescription (3.1) continues to hold, although in that case, we have

$$\eta^k = \int_{\Gamma_c} \mathbf{T}(\hat{\boldsymbol{\xi}}) \cdot \mathbf{n} \cdot \phi^k d\Gamma.$$

4. Application to drag reduction via discrete blow holes. Let us consider the situation where there are m subregions of Γ_c , say the set $\{\Gamma_k\}$, such that

$$\Gamma_c - \cup \Gamma_k \neq \emptyset, \quad \overline{\Gamma_k \cup \Gamma_j} = \emptyset \quad k \neq j.$$

Consider a set of vector-valued functions, $\{\phi^k\}$ in $\mathbf{H}^{1/2}(\Gamma_c)$ such that

$$\text{supp } \phi^k \subset \Gamma_k, \quad \forall k,$$

so, clearly, these functions are orthogonal; we also want them normalized in the $\mathbf{H}^{1/2}$ norm.

We will define our admissibility set for the control as

$$\mathcal{U} = \{ \mathbf{g} \in \mathbf{H}^{1/2}(\Gamma_c) : \mathbf{g} = \sum \gamma_k \phi^k, a_k \leq \gamma_k \leq b_k, a_k, b_k \in R, \quad \forall k \}.$$

Now we can apply the framework of the last section. Let T be identified with Euclidean m -space, with the usual basis; i.e.,

$$\{\mathbf{e}_1, \dots, \mathbf{e}_m\}.$$

The minimum principle is

$$\sum_k \int_{\Gamma_k} [\mathbf{T}(\hat{\mathbf{v}}) + \mathbf{T}(\boldsymbol{\xi})] \cdot \mathbf{n} \cdot (\mathbf{g} - \hat{\mathbf{g}}) d\Gamma \geq 0, \quad \forall \mathbf{g} \in \mathcal{U},$$

which reduces to

$$\sum_k \eta^k (g - \hat{g})_k \geq 0, \quad \forall \mathbf{g} = \{g_k\} \in \mathcal{U},$$

where

$$\eta^k = \int_{\Gamma_k} [\mathbf{T}(\hat{\mathbf{v}}) + \mathbf{T}(\hat{\boldsymbol{\xi}})] \cdot \mathbf{n} \cdot \boldsymbol{\phi}^k d\Gamma.$$

Note that an *apparent* difficulty in implementation is the normalization of the basis vectors in the $\frac{1}{2}$ norm. Suppose that, instead,

$$\int_{\Gamma} [(\phi_1^k)^2 + \dots + (\phi_n^k)^2] d\Gamma = 1, \quad \forall k.$$

so that they are actually an orthonormal set in $\mathbf{L}^2(\Gamma)$. But this is also the duality pairing, thus we can choose the $\boldsymbol{\psi}$'s to be the $\boldsymbol{\phi}$'s. But in this case,

$$\eta_k = \langle \mathbf{T}, \boldsymbol{\phi}^k \rangle_{-1/2} = \langle \mathbf{T}, \Lambda^{-1} \boldsymbol{\phi}^k \rangle$$

which means that the η 's would be difficult to compute. In any case, the question of computing the η^k is moot, unless we know whether or not property C is to hold. The question of efficiently computing the optimum is far from satisfactorily settled. The formula (3.1) gives us the structure of the optimal solution, but it may be prohibitively expensive to have to visit each vertex of the m -cube. Note that the $\frac{1}{2}$ norm of the basis functions is not necessarily difficult to calculate. A reasonable choice of basis functions, at least for a_k and b_k small, are the piecewise linear functions, i.e., hat functions over the holes. These $\mathbf{H}^{1/2}$ -norms can be computed. For example, if $n = 2$, and $\Gamma_1 = [0, b]$, and

$$\phi_1^1 = \begin{cases} (2h/b)x & \text{if } x \leq b/2 \\ 2h(1 - x/b) & \text{if } x > b/2 \end{cases}$$

then,

$$\|\phi_1^1\|_{1/2}^2 = h^2[b/6 + 16/b^2 + 8(\ln(2) - 1/4)].$$

Remark 1. Note that the vector (η^1, \dots, η^m) is matrix representation for a Jacobian of $J'(\hat{\mathbf{g}})$. Thus, if $\eta^j = 0$ then $\hat{g}^j \in [a_k, b_k]$ is arbitrary, and we have a one-parameter family of extremals, parametrized by $[a_j, b_j]$. In the application at hand, this means that the j th control hole could be removed without effect on the functional. It would be interesting to know whether this kind of degeneracy affects the convergence of the approximation.

Remark 2. Care must be taken to make the problem well-posed, not only in the sense of existence and regularity, but also with respect to the control. Degeneracies can arise in several ways: if property C does not hold then it is seen that the necessary conditions do not involve the cost functional; this means that the constraint set does not have the structure that we want; it is not a manifold, or solutions of the constraint equations are so sparse that a variational approach is not useful. On the other hand, if $\mathbf{g} \in \mathcal{U}$ supplies an unconstrained minimum (for example, if $\mathbf{v}_\infty = 0$ and $\mathbf{g} = 0$, then $J'(\hat{\mathbf{v}}) \equiv 0$, and the constraint is superfluous; in which case a result such as (3.1) would be useless. Note that we do not have to explicitly rule out this case, for then, as can be deduced from the previous remark, $\eta_k = 0$, for all k .

5. A simple computational example. Recall that the problem to be discretized is given by

$$\begin{aligned} a(\mathbf{v}, \mathbf{w}) + c(\mathbf{v}, \mathbf{v}, \mathbf{w}) + b(\mathbf{w}, p) &= 0, & \forall \mathbf{w} \in \mathbf{H}_0^1(\Omega) \\ b(\mathbf{v}, q) &= 0, & \forall q \in L_0^2(\Omega) \\ \mathbf{v}|_{\Gamma_k} &= \mathbf{g}_k \\ \mathbf{v}|_{\Gamma - (\cup \Gamma_k) \cup \Gamma_{(1)}} &= \mathbf{w}_0 \\ \mathbf{T}(\mathbf{v}) \cdot \mathbf{n}|_{\Gamma_{(2)}} &= 0, \end{aligned}$$

where \mathbf{w}_0 is a prescribed function. For the simple example given below we did not use the adjoint system in the updating procedure, only the theoretical fact that the optimal solution is bang-bang i.e., given by (3.1) and (3.2). The discretization is effected using a standard finite element algorithm (cf. [7]). Thus, we choose finite element subspaces

$$\mathbf{V}^h \subset \mathbf{H}^1(\Omega)$$

and

$$S^h \subset L_0^2(\Omega)$$

and require that $\mathbf{v}^h \in \mathbf{V}^h$ and $p^h \in S^h$ satisfy

$$\begin{aligned} a(\mathbf{v}^h, \mathbf{w}^h) + c(\mathbf{v}^h, \mathbf{v}^h, \mathbf{w}^h) + b(\mathbf{w}^h, p^h) &= 0, & \forall \mathbf{w}^h \in \mathbf{V}^h \\ b(\mathbf{v}^h, q^h) &= 0, & \forall q^h \in S^h \\ \mathbf{v}^h|_{\Gamma_k} &= \mathbf{g}_k^h, \\ \mathbf{v}^h|_{\Gamma - (\cup \Gamma_k) \cup \Gamma_{(1)}} &= \mathbf{w}_0^h, \end{aligned}$$

where \mathbf{g}_k^h and \mathbf{w}_0^h are suitable approximations to \mathbf{g}_k and \mathbf{w}_0 , respectively. (See the discussion above.)

The particular computational problem we consider here is described in Fig. 1.

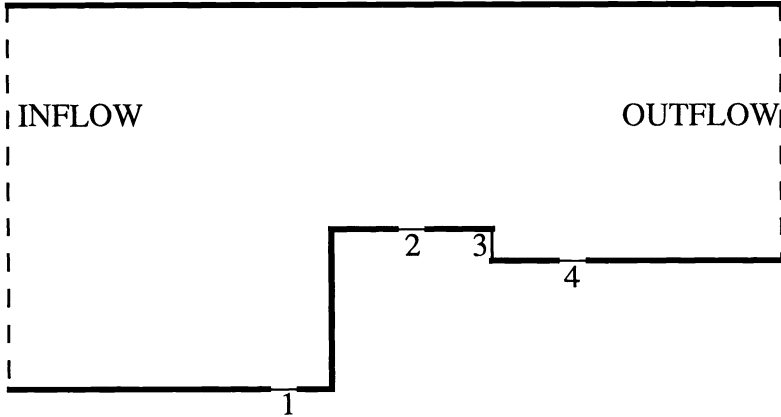


Fig. 1. Geometry and hole placement for flow over a forward facing step.

Here, Γ_k , $k = 1, 2, 3, 4$, are the regions indicated and \mathbf{w}_0 is zero on the top and bottom portions of $\Gamma - \cup_k \Gamma_k$ and nonzero on the inflow and outflow portions. This problem is of use in viscous drag reduction studies. The step represents a protuberance on a wing, e.g., due to struts, rivets, etc. These obstacles can trip separation or accelerate transition to turbulence. Our studies examine how the injection or suction of fluid can ameliorate the negative influence of the step. We use the Taylor-Hood element pair as amended by [12]. Continuous piecewise quadratic polynomials with respect to a triangulation of the flow domain are employed for the velocity components. For the pressure approximation, we use continuous piecewise linear polynomials with respect

to the same triangulation, augmented by a piecewise constant in each triangle. This locally mass conserving element pair is known to provide the error estimate (see [12])

$$\|\mathbf{v} - \mathbf{v}^h\|_0 + h(\|\mathbf{v} - \mathbf{v}^h\|_1 + \|p - p^h\|_0) \leq Ch^{s+1}(\|\mathbf{v}\|_{s+1} + \|p\|_s),$$

whenever

$$\mathbf{v} \in \mathbf{H}^{s+1}(\Omega), \quad p \in H^s(\Omega), \quad 1 \leq s \leq 2,$$

where h is a measure of the size of the finite element grid, and whenever \mathbf{g}_k^h and \mathbf{w}_0^h are accurate enough approximations. For example, the above estimate holds if we merely take \mathbf{g}_k^h and \mathbf{w}_0^h to be the interpolant, in \mathbf{V}^h restricted to the appropriate segment of Γ , of \mathbf{g}_k and \mathbf{w}_0 respectively.

For the computation we choose the velocity profile at the hole to satisfy

$$\mathbf{g}_k \cdot \mathbf{n} = \alpha_k p_{2k}, \quad \alpha_k \in [a_k, b_k] \quad \text{and} \quad \mathbf{g} \times \mathbf{n} = 0,$$

where p_{2k} is a quadratic polynomial vanishing at the edge of the hole. Then, our control set is determined by the appropriate parameters. From the theory presented above, we know that, for the optimal solution, α_k must be either a_k or b_k . Thus, to determine the optimal solution, we need to solve our finite element problem with \mathbf{g}_k given, following the above prescription. One may then compute $J(\mathbf{v})$ for each of these cases (16 in this example), and thus determine the minimum.

We choose the a_k, b_k so that the maximum mass flow in or out of any hole is 1/12 of the mass flow at the inflow. We compute the dissipation function with each hole blowing, sucking, or turned off. If only blowing is allowed, i.e., $a_k = 0$, it is found that the dissipation function is minimized if only the second hole is active. If only suction is allowed, i.e., $b_k = 0$, it is found that allowing only the third hole to be active is best. If both blowing and suction are allowed, i.e., $a_k = -b_k$, then having suction through the third hole and injection through the second hole is best.

REFERENCES

- [1] L. CATTABRIGA, *Su un problema al contorno relativo al sistema di equazione di Stokes*, Rend. Mat. Sem. Univ. Padova, 31 (1961), pp. 308–340.
- [2] G. DUVAUT AND J. LIONS, *Inequalities in Mechanics and Physics*, Springer, Berlin, 1976.
- [3] M. GUNZBURGER, L. HOU, AND T. SVOBODNY, *Analysis and finite element approximation of optimal control problems for the stationary Navier-Stokes equations with Dirichlet controls*, to appear in Math. Model. Numer. Anal.
- [4] G. HOUGH, *Viscous Flow Drag Reduction*, American Institute of Aeronautics and Astronautics, New York, 1980.
- [5] A. IOFFE AND V. TIKHOMOROV, *Extremal Problems*, North-Holland, Amsterdam, 1979.
- [6] V. GIRAUT AND P.-A. RAVIART, *Finite Element Approximation of the Navier-Stokes Equations*, Springer, Berlin, 1979.
- [7] V. GIRAUT AND P.-A. RAVIART, *Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms*, Springer, New York, 1986.
- [8] O. LADYZHENSKAYA, *The Mathematical Theory of Viscous Incompressible Flow*, Gordon and Breach, New York, 1963.
- [9] J. SERRIN, *Mathematical principles of classical fluid mechanics*, Handbuch der Physik, VIII/1 (1959), pp. 1–125.
- [10] R. TEMAM, *Navier-Stokes Equations*, North-Holland, New York, 1979.
- [11] R. TEMAM, *Navier-Stokes Equations and Nonlinear Functional Analysis*, SIAM, Philadelphia, 1983.
- [12] R. THATCHER, *Locally mass-conserving Taylor-Hood elements for two- and three-dimensional flow*, Inter. J. Numer. Meth. Fluids, 11 (1990), pp. 341–353.
- [13] V. TIKHOMOROV, *Fundamental Principles of the Theory of Extremal Problems*, Wiley, Chichester, 1982.

THE DISCRETE TIME H_∞ CONTROL PROBLEM WITH MEASUREMENT FEEDBACK*

A. A. STOORVOGEL†

Abstract. This paper is concerned with the discrete time H_∞ control problem with measurement feedback. It follows that, as in the continuous time case, the existence of an internally stabilizing controller that makes the H_∞ norm strictly less than 1 is related to the existence of stabilizing solutions to two algebraic Riccati equations. However, in the discrete time case, the solutions of these algebraic Riccati equations must satisfy extra conditions.

Key words. H_∞ control, discrete time, algebraic Riccati equation, measurement feedback.

AMS(MOS) subject classifications. 93C05, 93C35, 93C45, 93C55, 49B99, 49C05

1. Introduction. The H_∞ control problem with measurement feedback has been thoroughly investigated (see, e.g., [5], [6], [7], [13], [14], [21], [22], [24], [28]). However, all of these papers discuss the continuous time case. In this paper, in contrast with the above papers, we discuss the discrete time case.

In practical applications, most people are mainly concerned with discrete time systems. One major reason is that to control a continuous time system we often apply a digital computer on which we can only implement a discrete time controller. One possible approach is to derive a continuous time H_∞ controller and then to discretize the controller so that a computer may be used.

Discretizing the system first and then using H_∞ control designed for discrete time systems might be a more useful approach. This comparison can, only be made, however, after the discrete time H_∞ control problem has been solved. Taking the effects of discretization into account is another possibility, see [3], [4].

Also, certain systems are in themselves inherently discrete, and certainly for these systems it is useful to have results available for H_∞ control problems.

One approach is to apply a transformation in the frequency-domain that transforms discrete time systems to continuous time systems. The transformation we have in mind is discussed, for instance, in [8, App. 1]. With this transformation, discrete time H_∞ functions are mapped isometrically onto continuous time H_∞ functions. We can then use the results available for continuous time systems and afterward apply the inverse transformation on the controller thus obtained.

This transformation, however, is not always attractive. It maps systems with a pole in 1 into nonproper systems. Also this transformation is such that it clouds the understanding of specific features of discrete time H_∞ control because of this complex transformation. If it is possible to derive results for discrete time systems, why not apply these results directly instead of performing this unnatural transformation?

In the papers on H_∞ control with continuous time, several methods were used to solve the H_∞ control problem. Recently, a paper solving the discrete time H_∞ control problem using frequency domain techniques has appeared (see [12]). Also the polynomial approach has been applied to discrete time systems (see [11]). Derivation of the results for the discrete time H_∞ control problem could probably also be based on the work of [26]. In addition, several papers have appeared using a time-domain approach

* Received by the editors November 27, 1989; accepted for publications (in revised form) December 11, 1990.

† Department of Mathematics and Computing science, Eindhoven University of Technology, Post Office Box 513, 5600 MB Eindhoven, the Netherlands.

(see [1], [16], [27]). However, [16] does not contain any proof of the results obtained, and [1], [27] make a number of extra assumptions on the system under consideration. In [1], [16], [27] the authors first investigate the finite horizon problem and then derive a solution of the infinite horizon problem by considering it as a kind of limiting case as the endpoint tends to infinity.

In contrast, this paper directly investigates the infinite horizon case. We use time-domain techniques that have many familiarities with those used in [22], [24], which deal with the continuous time case. The method used in this paper was derived independently from [1], [16], [27]. The current paper is an extension of [23], which deals with the full-information case. However, contrary to the latter paper, here we give detailed proofs of all our results.

We assume that two particular transfer matrices are left- and right-invertible, respectively. The only other assumption we must make is that two subsystems have no invariant zeros on the unit circle. Our assumptions are exactly the discrete time analogues of the assumptions in [9]. The assumptions we make are weaker than the assumptions in [12], [27], and the same as the ones made in [16].

As in the continuous time case, the necessary and sufficient conditions for the existence of suitable controllers involve positive semidefinite stabilizing solutions of two algebraic Riccati equations. As in the continuous time case, the quadratic term in these algebraic Riccati equations is indefinite. However, compared to the continuous time case, the solutions of these equations must satisfy another assumption: matrices depending on these solutions should be positive definite.

The outline of this paper is as follows. In §2 we will formulate the problem and give our main results. In §3 we will derive the existence of a stabilizing solution of the first algebraic Riccati equation starting from the assumption that there exists an internally stabilizing feedback that makes the H_∞ norm of the closed loop system less than 1. In §4 we will show the existence of a stabilizing solution of the second algebraic Riccati equation and complete the proof that our conditions are necessary. This is done by transforming the original system into a new system with the property that a controller “works” for the new system if and only if it “works” for the original system. In §5 it is shown that our conditions are also sufficient. It follows that the system transformation of §4 repeated in a dual form exactly gives the desired results. We will end with some concluding remarks in §6.

2. Problem formulation and main results. By \mathcal{N} and \mathcal{R} we denote the natural numbers and the real numbers, respectively. Moreover, by σ we denote the shift $(\sigma x)(k) := x(k+1)$ for all $k \in \mathcal{N}$. At a certain stage, we also need a backward difference equation of the form $\sigma^{-1}x = Ax + Bu$. We define the solution x to be a mapping from $\mathcal{N} \cup \{-1\}$ to \mathcal{R}^n given by

$$\begin{cases} x|_{\mathcal{N}} = \sigma A(x|_{\mathcal{N}}) + \sigma Bu, \\ x(-1) = Ax(0) + Bu(0). \end{cases}$$

It will follow that extending this function from \mathcal{N} to $\mathcal{N} \cup \{-1\}$ is a useful definition.

We consider the following time-invariant system:

$$(2.1) \quad \Sigma : \begin{cases} \sigma x = Ax + Bu + Ew, \\ y = C_1x + \quad \quad + D_{12}w, \\ z = C_2x + D_{21}u + D_{22}w, \end{cases}$$

where for all $k \in \mathcal{N}$, $x(k) \in \mathcal{R}^n$ is the state, $u(k) \in \mathcal{R}^m$ is the control input, $y(k) \in \mathcal{R}^l$ is the measurement, $w(k) \in \mathcal{R}^9$ is the unknown disturbance, and $z(k) \in \mathcal{R}^p$ is the

output to be controlled. $A, B, E, C_1, C_2, D_{12}, D_{21}$, and D_{22} are matrices of appropriate dimension.

If we apply a dynamic feedback law $u = Fy$ to Σ , then the closed loop system with zero initial conditions defines a convolution operator $\Sigma_{cl,F}$ from w to y . We seek a feedback law $u = Fy$ that is internally stabilizing and that minimizes the ℓ_2 -induced operator norm of $\Sigma_{cl,F}$ over all internally stabilizing feedback laws. We will investigate dynamic feedback laws of the form

$$(2.2) \quad \Sigma_F : \begin{cases} \sigma p = Kp + Ly, \\ u = Mp + Ny. \end{cases}$$

We will say that the dynamic compensator Σ_F , given by (2.2), is internally stabilizing when applied to the system Σ , described by (2.1), if the following matrix is asymptotically stable:

$$(2.3) \quad \begin{pmatrix} A + BNC_1 & BM \\ LC_1 & K \end{pmatrix};$$

i.e., all its eigenvalues lie in the open unit disc. Denote by G_F the closed loop transfer matrix. The ℓ_2 -induced operator norm of the convolution operator $\Sigma_{cl,F}$ is equal to the H_∞ norm of the transfer matrix G_F and is given by

$$\begin{aligned} \|G_F\|_\infty &:= \sup_{\theta \in [0, 2\pi]} \|G_F(e^{i\theta})\| \\ &= \sup_w \left\{ \frac{\|z\|_2}{\|w\|_2} \mid w \in \ell_2^l, w \neq 0 \right\}, \end{aligned}$$

where the ℓ_2 -norm is given by

$$\|p\|_2 := \left(\sum_{k=0}^{\infty} p^T(k)p(k) \right)^{1/2}$$

and where $\|\cdot\|$ denotes the largest singular value. We refer to this norm as the H_∞ norm of the closed loop system.

In this paper we will derive necessary and sufficient conditions for the existence of a dynamic compensator Σ_F that is internally stabilizing and which is such that the closed loop transfer matrix G_F satisfies $\|G_F\|_\infty < 1$. Furthermore, if a stabilizing Σ_F exists, which makes the H_∞ norm of the closed loop system less than 1, then we derive an explicit formula for one particular Σ_F satisfying these requirements.

By scaling the plant, we can thus, in principle, find the infimum of the H_∞ norm of the closed loop system over all stabilizing controllers. This will involve a search procedure.

In the formulation of our main result we will need the concept of invariant zero: z_0 is called an *invariant zero* of the system (A, B, C, D) if

$$\text{rank}_{\mathcal{R}} \begin{pmatrix} z_0 I - A & -B \\ C & D \end{pmatrix} < \text{rank}_{\mathcal{R}(z)} \begin{pmatrix} zI - A & -B \\ C & D \end{pmatrix},$$

where $\text{rank}_{\mathcal{K}}$ denotes the rank as a matrix with entries in the field \mathcal{K} . By $\mathcal{R}(z)$ we denote the field of real rational functions. The system (A, B, C, D) is called *left- (right-) invertible* if the transfer matrix $C(zI - A)^{-1}B + D$ is left- (right-) invertible

as a matrix with entries in the field of real rational functions. We can now formulate our main result.

THEOREM 2.1. *Consider system (2.1). Assume that the system (A, B, C_2, D_{21}) has no invariant zeros on the unit circle and is left-invertible. Moreover, assume that the system (A, E, C_1, D_{12}) has no invariant zeros on the unit circle and is right invertible. The following statements are equivalent:*

- (i) *There exists a dynamic compensator Σ_F of the form (2.2) such that the resulting closed loop transfer matrix G_F satisfies $\|G_F\|_\infty < 1$ and the closed loop system is internally stable.*
- (ii) *There exist symmetric matrices $P \geq 0$ and $Y \geq 0$ such that*
 - (a) *We have*

$$(2.4) \quad V > 0, \quad R > 0,$$

where

$$\begin{aligned} V &:= B^T P B + D_{21}^T D_{21}, \\ R &:= I - D_{22}^T D_{22} - E^T P E \\ &\quad + (E^T P B + D_{22}^T D_{21}) V^{-1} (B^T P E + D_{21}^T D_{22}). \end{aligned}$$

This implies that the matrix $G(P)$ is invertible, where

$$(2.5) \quad G(P) := \begin{pmatrix} D_{21}^T D_{21} & D_{21}^T D_{22} \\ D_{22}^T D_{21} & D_{22}^T D_{22} - I \end{pmatrix} + \begin{pmatrix} B^T \\ E^T \end{pmatrix} P \begin{pmatrix} B & E \end{pmatrix}.$$

- (b) *P satisfies the discrete algebraic Riccati equation*

$$(2.6) \quad P = A^T P A + C_2^T C_2 - \begin{pmatrix} B^T P A + D_{21}^T C_2 \\ E^T P A + D_{22}^T C_2 \end{pmatrix}^T G(P)^{-1} \begin{pmatrix} B^T P A + D_{21}^T C_2 \\ E^T P A + D_{22}^T C_2 \end{pmatrix}.$$

- (c) *The matrix $A_{cl,P}$ is asymptotically stable, where*

$$(2.7) \quad A_{cl,P} := A - \begin{pmatrix} B & E \end{pmatrix} G(P)^{-1} \begin{pmatrix} B^T P A + D_{21}^T C_2 \\ E^T P A + D_{22}^T C_2 \end{pmatrix}.$$

Moreover, if, given the matrix P satisfying (a)–(c), we define the following matrices:

$$\begin{aligned} H &:= E^T P A + D_{22}^T C_2 - [E^T P B + D_{22}^T D_{21}] V^{-1} [B^T P A + D_{21}^T C_2], \\ A_P &:= A + E R^{-1} H, \\ E_P &:= E R^{-1/2}, \\ C_{1,P} &:= C_1 + D_{12} R^{-1} H, \\ C_{2,P} &:= V^{-1/2} (B^T P A + D_{21}^T C_2) + V^{-1/2} [B^T P E + D_{21}^T D_{22}] R^{-1} H, \\ D_{12,P} &:= D_{12} R^{-1/2}, \\ D_{21,P} &:= V^{1/2}, \\ D_{22,P} &:= V^{-1/2} (B^T P E + D_{21}^T D_{22}) R^{-1/2}, \end{aligned}$$

then the matrix Y should satisfy

(d) *We have*

$$(2.8) \quad W > 0, \quad S > 0,$$

where

$$\begin{aligned} W &:= D_{12,P} D_{12,P}^T + C_{1,P} Y C_{1,P}^T, \\ S &:= I - D_{22,P} D_{22,P}^T - C_{2,P} Y C_{2,P}^T \\ &\quad + (C_{2,P} Y C_{1,P}^T + D_{22,P} D_{12,P}^T) W^{-1} (C_{1,P} Y C_{2,P}^T + D_{12,P} D_{22,P}^T). \end{aligned}$$

This implies that the matrix $H_P(Y)$ is invertible, where

$$(2.9) \quad H_P(Y) := \begin{pmatrix} D_{12,P} D_{12,P}^T & D_{12,P} D_{22,P}^T \\ D_{22,P} D_{12,P}^T & D_{22,P} D_{22,P}^T - I \end{pmatrix} + \begin{pmatrix} C_{1,P} \\ C_{2,P} \end{pmatrix} Y \begin{pmatrix} C_{1,P} \\ C_{2,P} \end{pmatrix}^T.$$

(e) *Y satisfies the following discrete algebraic Riccati equation:*

$$(2.10) \quad Y = A_P Y A_P^T + E_P E_P^T - \begin{pmatrix} C_{1,P} Y A_P^T + D_{12,P} E_P^T \\ C_{2,P} Y A_P^T + D_{22,P} E_P^T \end{pmatrix}^T H_P(Y)^{-1} \begin{pmatrix} C_{1,P} Y A_P^T + D_{12,P} E_P^T \\ C_{2,P} Y A_P^T + D_{22,P} E_P^T \end{pmatrix}.$$

(f) *The matrix $A_{cl,P,Y}$ is asymptotically stable, where*

$$(2.11) \quad A_{cl,P,Y} := A_P - \begin{pmatrix} C_{1,P} Y A_P^T + D_{12,P} E_P^T \\ C_{2,P} Y A_P^T + D_{22,P} E_P^T \end{pmatrix}^T H_P(Y)^{-1} \begin{pmatrix} C_{1,P} \\ C_{2,P} \end{pmatrix}.$$

In the case where there exist $P \geq 0$ and $Y \geq 0$ satisfying (ii), then a controller of the form (2.2) satisfying the requirements in (i) is given by

$$\begin{aligned} N &:= -D_{21,P}^{-1} (C_{2,P} Y C_{1,P}^T + D_{22,P} D_{12,P}^T) W^{-1}, \\ M &:= -(D_{21,P}^{-1} C_{2,P} + N C_{1,P}), \\ L &:= B N + (A_P Y C_{1,P}^T + E_P D_{12,P}^T) W^{-1}, \\ K &:= A_{cl,P} - L C_{1,P}. \end{aligned}$$

Remarks.

(i) Necessary and sufficient conditions for the existence of an internally stabilizing feedback compensator, which makes the H_∞ norm of the closed loop system less than some, a priori given, upper bound $\gamma > 0$, can be easily derived from Theorem 2.1 by scaling.

(ii) If we compare these conditions with the conditions for the continuous time case (see [6], [22]) we note that conditions (2.4) and (2.8) are this time depending on the solutions of the two Riccati equations. A simple example showing that the assumption $G(P)$ invertible is not sufficient is given by the system

$$\begin{cases} \sigma x &= u + 2w, \\ y &= \begin{pmatrix} 1 \\ 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ 1 \end{pmatrix} w, \\ z &= \begin{pmatrix} 1 \\ 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u. \end{cases}$$

There does not exist a dynamic compensator satisfying the requirements of part (i) of Theorem 2.1, but there does exist a positive semidefinite matrix P satisfying (2.6) such that matrix (2.7) is asymptotically stable, namely $P = 1$. However, for this P we have $R = -1$. Therefore matrices like E_p are ill-defined and we cannot even look for a matrix Y satisfying (2.8)–(2.11).

(iii) Since our starting point of the proof of (i) \Rightarrow (ii) will not be part (i) of theorem 2.1 but Condition 3.2, it can be shown that we cannot make the H_∞ norm less by allowing more general, possibly even nonlinear, causal feedbacks.

The proof of the existence of a stabilizing solution of the Riccati equation will be reminiscent of the proof given in [24] for the continuous time case. However due to our weaker assumptions and conditions (2.4) and (2.8), there are quite a number of extra intricacies. The remainder of the proof is based on [22].

Another interesting case was discussed in [23]. However, the latter reference only gives the general outline of the proof. In contrast, the present paper will give much more detail. Reference [23] discusses the so-called full-information case, shown below.

Full information case. $C_1 = \begin{pmatrix} I \\ 0 \end{pmatrix}$, $D_{12} = \begin{pmatrix} 0 \\ I \end{pmatrix}$.

In this case, we have $y_1 = x$ and $y_2 = w$; i.e., we know both the state and the disturbance of the system at time k . However, we cannot apply Theorem 2.1 to this case since system (A, E, C_1, D_{12}) is not right-invertible. Nevertheless, following the proof for this special case, it can be shown that there exists a feedback satisfying part (i) of Theorem 2.1 if and only if there exists a symmetric matrix $P \geq 0$ satisfying conditions (a)–(c) of part (ii) of Theorem 2.1. Moreover, in that case we can find static output feedbacks $u = F_1 x + F_2 w$ with the desired properties. One particular choice for $F = (F_1, F_2)$ is given by

$$(2.12) \quad F_1 := -(D_{21}^T D_{21} + B^T P B)^{-1} (B^T P A + D_{21}^T C_2),$$

$$(2.13) \quad F_2 := -(D_{21}^T D_{21} + B^T P B)^{-1} (B^T P E + D_{21}^T D_{22}).$$

3. Existence of stabilizing solutions of the Riccati equations. In this section we assume that part (i) of Theorem 2.1 is satisfied. We will show that the existence of P satisfying conditions parts (a)–(c) in (ii) is necessary.

Consider system (2.1). For given disturbance w and control input u let $x_{u,w,\xi}$ and $z_{u,w,\xi}$ denote, respectively, the resulting state and output for initial state $x(0) = \xi$. If $\xi = 0$ we will simply write $x_{u,w}$ and $z_{u,w}$. We first give a definition.

DEFINITION 3.1. An operator $f : \ell_2 \rightarrow \ell_2$, $w \rightarrow f(w)$ is called causal if for any $w_1, w_2 \in \ell_2$, and $k \in \mathcal{N}$:

$$w_1|_{[0,k]} = w_2|_{[0,k]} \Rightarrow f(w_1)|_{[0,k]} = f(w_2)|_{[0,k]}.$$

f is called strictly causal if for any $w_1, w_2 \in \ell_2$, and $k \in \mathcal{N}$ we have

$$w_1|_{[0,k-1]} = w_2|_{[0,k-1]} \Rightarrow f(w_1)|_{[0,k]} = f(w_2)|_{[0,k]}.$$

A controller of the form (2.2) always defines a causal operator. In the case where $N = 0$, this operator is strictly causal. We will label the following condition.

CONDITION 3.2. (A, B) stabilizable and for system (2.1) there exists causal $f : \ell_2^l \rightarrow \ell_2^n$ and $\delta < 1$ such that for all $w \in \ell_2^l$ with $u = f(w)$ we have $x_{u,w} \in \ell_2^n$ and $\|z_{u,w}\|_2 \leq \delta \|w\|_2$.

If there exists a dynamic compensator Σ_F such that $\|G_F\|_\infty < 1$ and such that the closed loop system is internally stable, then Condition 3.2 is satisfied. Hence if the

requirements of part (i) of Theorem 2.1 are satisfied, then Condition 3.2 holds. Note that Condition 3.2 is equivalent to the requirement that there exists a causal operator f such that the feedback $u = f(x, w)$ satisfies Condition 3.2. This follows from the fact that, after applying the feedback, there exists a causal operator g mapping w to x and, therefore, we could have started with the causal operator $u = f(g(w), w)$ in the first place. Conversely, if we have the feedback $u = f(w)$, then we define $f_1(x, w) := f(w)$, which then satisfies the requirements of the reformulated Condition 3.2.

Finally, we would like to remark that besides the obvious condition that (A, B) should be stabilizable, there is a more implicit extra condition $x_{u,w} \in \ell_2^n$, which is also related to stability. Intuitively, these two conditions imply that we cannot only find a controller that is input-output stabilizing (i.e., the closed loop transfer matrix is in H_∞) and that makes the H_∞ of the closed loop system less than 1, but even an internally stabilizing controller with the same property. This is only true for the full-information case (see [23]). For the more general measurement-feedback case, we need extra stability conditions related to detectability.

We will show that the existence of such causal f and $\delta < 1$ satisfying Condition 3.2 already implies that there exists a positive semidefinite solution of the discrete algebraic Riccati equation (2.6) such that (2.7) is asymptotically stable and (2.4) is satisfied. *We will assume for the time being that*

$$(3.1) \quad D_{21}^T [C_2 \quad D_{22}] = 0,$$

and we will derive the more general statement later. To prove the existence of the desired P , we will investigate the following sup-inf problem:

$$(3.2) \quad \mathcal{C}^*(\xi) := \sup_{w \in \ell_2^n} \inf_u \{ \|z_{u,w,\xi}\|_2^2 - \|w\|_2^2 \mid u \in \ell_2^m \text{ such that } x_{u,w,\xi} \in \ell_2^n \}$$

for arbitrary initial state ξ . We will prove that Condition 3.2 implies that $\mathcal{C}^*(\xi)$ is finite for every ξ . Moreover, it will be shown that there exists a $P \geq 0$ such that $\mathcal{C}^*(\xi) = \xi^T P \xi$. At the end of this section, we then prove that this P exactly satisfies conditions (a)–(c) of Theorem 2.1. We first infimize, for given $w \in \ell_2$ and $\xi \in \mathcal{R}^n$, the function $\|z_{u,w,\xi}\|_2^2 - \|w\|_2^2$ over all $u \in \ell_2$ for which $x_{u,w,\xi} \in \ell_2$. After that, we maximize over $w \in \ell_2$.

Our proof is based on Pontryagin's maximum principle. This only gives necessary conditions for optimality. However, in [15] a sufficient condition for optimality is derived over a finite horizon. We will use the ideas from [15], together with our stability requirement, $x_{u,w,\xi} \in \ell_2$, to adapt the proof to the infinite horizon case.

We start by constructing a solution of the adjoint Hamilton–Jacobi equation, which is a natural starting point if we use Pontryagin's maximum principle.

Let L be such that $D_{21}^T D_{21} + B^T L B$ is invertible and such that L is the positive semidefinite solution of the following discrete algebraic Riccati equation:

$$(3.3) \quad L = A^T L A + C_2^T C_2 - A^T L B (D_{21}^T D_{21} + B^T L B)^{-1} B^T L A$$

for which

$$(3.4) \quad A_L := A - B (D_{21}^T D_{21} + B^T L B)^{-1} B^T L A$$

is asymptotically stable. The existence of such L is guaranteed under the assumption that (A, B, C_2, D_{21}) has no invariant zeros on the unit circle and is left-invertible and,

moreover, that (A, B) is stabilizable (see [20]). We define

$$(3.5) \quad r(k) := - \sum_{i=k}^{\infty} [X_1 A^T]^{i-k} X_1 (LEw(i) + C_2^T D_{22}w(i+1)),$$

where

$$(3.6) \quad X_1 := I - LB(D_{21}^T D_{21} + B^T LB)^{-1} B^T.$$

Note that r is well defined since the matrix $A_L = X_1^T A$ is asymptotically stable, which implies that $X_1 A^T$ is asymptotically stable. Next, we define the functions y, \tilde{x} , and η by

$$(3.7) \quad y := M^{-1} B^T [A^T \sigma r - LEw - C_2^T D_{22} \sigma w],$$

$$(3.8) \quad \sigma \tilde{x} = A_L \tilde{x} + By + Ew, \quad \tilde{x}(0) = \xi,$$

$$(3.9) \quad \eta := -X_1 LA \tilde{x} + r,$$

where $M := D_{21}^T D_{21} + B^T LB$. Since $X_1 A^T$ is asymptotically stable, it can be checked straightforwardly that, given $\xi \in \mathcal{R}^n$ and $w \in \ell_2^l$, we have $r, \tilde{x}, \eta \in \ell_2$.

After some standard calculations, we find the following lemma.

LEMMA 3.3. *Let $\xi \in \mathcal{R}^n$ and $w \in \ell_2^l$ be given. The function $\eta \in \ell_2^n$ is a solution of the following backward difference equation:*

$$(3.10) \quad \sigma^{-1} \eta = A^T \eta - C_2^T C_2 \tilde{x} - C_2^T D_{22} w, \quad \lim_{k \rightarrow \infty} \eta(k) = 0.$$

Here η is extended to a function from $\mathcal{N} \cup \{-1\}$ to \mathcal{R}^n by choosing $\eta(-1)$ such that (3.10) is satisfied.

In the statement of Pontryagin's maximum principle, this equation is the so-called "adjoint Hamilton-Jacobi equation," and η is called the "adjoint state variable." We have constructed a solution to this equation and we show that this η indeed yields a minimizing u . Note the difference with the continuous time case (see [24]), where we could derive a differential equation forward in time, while in discrete time we can only derive a difference equation forward in time when A is invertible. To prevent these kinds of difficulties, it is assumed in [12] that A is invertible. The following lemma states that η yields a minimizing u .

LEMMA 3.4. *Let system (2.1) be given. Moreover, let w and ξ be fixed. Then*

$$\begin{aligned} \tilde{u} &:= -(D_{21}^T D_{21} + B^T LB)^{-1} B^T LA \tilde{x} + y \\ &= \arg \inf_u \{ \|z_{u,w,\xi}\|_2 \mid u \in \ell_2^m \text{ such that } x_{u,w,\xi} \in \ell_2^n \}. \end{aligned}$$

Proof. It can be easily checked that $\tilde{x} = x_{\tilde{u},w,\xi}$. Define

$$\mathcal{J}_T(u) := \sum_{i=0}^T \|C_2 x_{u,w,\xi}(i) + D_{21} u(i) + D_{22} w(i)\|^2.$$

Let $u \in \ell_2^m$ be an arbitrary control input such that $x_{u,w,\xi} \in \ell_2^n$. We find that

$$\begin{aligned} &\mathcal{J}_T(u) - \mathcal{J}_{T-1}(u) - 2\eta^T(T)x(T+1) + 2\eta^T(T-1)x(T) \\ &\quad \|C_2 x(T)\|^2 + [D_{21}^T D_{21} u(T) - 2B^T \eta(T)]^T u(T) - 2\eta^T(T)Ew(T) - 2x^T(T)C_2^T C_2 \tilde{x}(T). \end{aligned}$$

We also find that

$$\begin{aligned} \mathcal{J}_T(\tilde{u}) - \mathcal{J}_{T-1}(\tilde{u}) - 2\eta^T(T)\tilde{x}(T+1) + 2\eta^T(T-1)\tilde{x}(T) \\ = -\|C_2\tilde{x}(T)\|^2 + [D_{21}^T D_{21}\tilde{u}(T) - 2B^T\eta(T)]^T \tilde{u}(T) - 2\eta^T(T)Ew(T). \end{aligned}$$

Hence if we sum the last two equations from zero to infinity and subtract from each other we find that

$$\begin{aligned} \|z_{\tilde{u},w,\xi}\|_2^2 - \|z_{u,w,\xi}\|_2^2 &= \sum_{i=0}^{\infty} -\|C_2(x(i) - \tilde{x}(i))\|^2 \\ &+ \sum_{i=0}^{\infty} [D_{21}^T D_{21}\tilde{u}(i) - 2B^T\eta(i)]^T \tilde{u}(i) - [D_{21}^T D_{21}u(i) - 2B^T\eta(i)]^T u(i). \end{aligned}$$

It can easily be checked that $B^T\eta(i) = D_{21}^T D_{21}\tilde{u}(i)$ for all i . Therefore, for every i we have

$$[D_{21}^T D_{21}\tilde{u}(i) - 2B^T\eta(i)]^T \tilde{u}(i) = \inf_u [D_{21}^T D_{21}u - 2B^T\eta(i)]^T u.$$

Together, the last two equations imply that $\|z_{\tilde{u},w,\xi}\|_2 \leq \|z_{u,w,\xi}\|_2$, which is exactly what we had to prove. Since (A, B, C_2, D_{21}) is left-invertible, it can easily be shown that the minimizing u is unique. \square

We now maximize over $w \in \ell_2^l$. This will then yield $\mathcal{C}^*(\xi)$. Define $\mathcal{F}(\xi, w) := (\tilde{x}, \tilde{u}, \eta)$ and $\mathcal{G}(\xi, w) := z_{\tilde{u},w,\xi} = C_2\tilde{x} + D_{21}\tilde{u} + D_{22}w$. It is clear from the previous lemma that \mathcal{F} and \mathcal{G} are bounded linear operators. Define

$$\begin{aligned} \mathcal{C}(\xi, w) &:= \|\mathcal{G}(\xi, w)\|_2^2 - \|w\|_2^2, \\ \|w\|_C &:= (-\mathcal{C}(0, w))^{1/2}. \end{aligned}$$

It can be easily shown that $\|\cdot\|_C$ defines a norm on ℓ_2^l . Using Condition 3.2, it can be shown straightforwardly that

$$(3.11) \quad \|w\|_2 \geq \|w\|_C \geq \rho\|w\|_2,$$

where $\rho > 0$ is such that $\rho^2 = 1 - \delta^2$ and δ is such that Condition 3.2 is satisfied. Hence $\|\cdot\|_C$ and $\|\cdot\|_2$ are equivalent norms.

Note that Lemma 3.4 still holds if Condition 3.2 does not hold. However, the result that $\|\cdot\|_C$ is a norm and that even $\|\cdot\|_C$ and $\|\cdot\|_2$ are equivalent norms is the essential property, which is implied by Condition 3.2 and which is the key to our derivation.

We have

$$(3.12) \quad \mathcal{C}^*(\xi) = \sup_{w \in \ell_2^l} \mathcal{C}(\xi, w).$$

We can derive the following properties of \mathcal{C}^* .

LEMMA 3.5.

(i) For all $\xi \in \mathcal{R}^n$ we have

$$(3.13) \quad 0 \leq \xi^T L \xi \leq \mathcal{C}^*(\xi) \leq \frac{\xi^T L \xi}{1 - \delta^2},$$

where δ is such that Condition 3.2 is satisfied.

(ii) For all $\xi \in \mathcal{R}^n$ there exists a unique $w_* \in \ell_2^l$ such that $\mathcal{C}^*(\xi) = \mathcal{C}(\xi, w_*)$.

Proof. Part (i). It is well known that L , as the stabilizing solution of the discrete time algebraic Riccati equation (3.3), is the cost of the discrete time, linear quadratic problem with internal stability (see [20]). Hence $\|\mathcal{G}(\xi, 0)\|_2^2 = \mathcal{C}(\xi, 0) = \xi^T L \xi$. Therefore we have $0 \leq \xi^T L \xi \leq \mathcal{C}^*(\xi)$. Moreover,

$$\begin{aligned} \mathcal{C}(\xi, w) &= \|\mathcal{G}(\xi, w)\|_2^2 - \|w\|_2^2 \\ &\leq (\|\mathcal{G}(\xi, 0)\|_2 + \|\mathcal{G}(0, w)\|_2)^2 - \|w\|_2^2 \\ &\leq \left(\sqrt{\xi^T L \xi} + \delta \|w\|_2 \right)^2 - \|w\|_2^2 \\ &\leq \frac{\xi^T L \xi}{1 - \delta^2}. \end{aligned}$$

Part (ii) can be proved in the same way as in [24]. First, show that $\|\cdot\|_C$ satisfies

$$(3.14) \quad -\|w_\alpha - w_\beta\|_C^2 = 2\mathcal{C}(\xi, w_\alpha) + 2\mathcal{C}(\xi, w_\beta) - 4\mathcal{C}\left(\xi, \frac{1}{2}(w_\alpha + w_\beta)\right)$$

for arbitrary $\xi \in \mathcal{R}^n$. Then it can be shown that a maximizing sequence of $\mathcal{C}(\xi, w)$ is a Cauchy sequence with respect to the $\|\cdot\|_C$ -norm and hence, since $\|\cdot\|_C$ and $\|\cdot\|_2$ are equivalent norms, there exists a maximizing ℓ_2 function w_* . It is easy to show uniqueness using (3.14). \square

Define $\mathcal{H} : \mathcal{R}^n \rightarrow \ell_2^l$, $\xi \rightarrow w_*$. Unlike the explicit expression for \tilde{u} we can only derive an implicit formula for w_* . However, we show with the following lemma that w_* is the unique solution of a linear equation.

LEMMA 3.6. *Let $\xi \in \mathcal{R}^n$ be given. Then $w_* = \mathcal{H}\xi$ is the unique ℓ_2 -function w satisfying:*

$$(3.15) \quad (I - D_{22}^T D_{22})w = -E^T \eta + D_{22}^T C_2 x,$$

where $(x, u, \eta) = \mathcal{F}(\xi, w)$.

Proof. Define $(x_*, u_*, \eta_*) = \mathcal{F}(\xi, w_*)$. Moreover, define $w_0 := -E^T \eta(w_*) + D_{22}^T D_{22} w_* + D_{22}^T C_2 x_*$ and $(x_0, u_0, \eta_0) := \mathcal{F}(\xi, w_0)$. We find that

$$(3.16) \quad \begin{aligned} &\|z_{u_0, w_0, \xi}(T)\|^2 - \|w_0(T)\|^2 - 2\eta_*^T(T)x_0(T+1) + 2\eta_*^T(T-1)x_0(T) \\ &= \|z_{u_*, w_*, \xi}(T) - z_{u_0, w_0, \xi}(T)\|^2 - \|z_{u_*, w_*, \xi}(T)\|^2 + \|w_0(T)\|^2 \end{aligned}$$

Here we use the fact that $D_{21}^T D_{21} u_*(i) = B^T \eta_*(i)$ for all i . We also find that

$$(3.17) \quad \begin{aligned} &\|z_{u_*, w_*, \xi}(T)\|^2 - \|w_*(T)\|^2 - 2\eta_*^T(T)x_*(T+1) + 2\eta_*^T(T-1)x_*(T) \\ &= 2w_0^T(T)w_*(T) - \|z_{u_*, w_*, \xi}(T)\|^2 - \|w_*(T)\|^2 \end{aligned}$$

Summing (3.16) and (3.17) from zero to infinity and subtracting from each other gives us

$$(3.18) \quad \mathcal{C}(\xi, w_*) = \mathcal{C}(\xi, w_0) - \|w_0 - w_*\|_2^2 - \|z_{u_0, w_0, \xi} - z_{u_*, w_*, \xi}\|_2^2.$$

Since w_* maximizes $\mathcal{C}(\xi, w)$ over all w , this implies $w_0 = w_*$.

That w_* is the unique solution of the equation (3.15) can be shown in a similar way. Assume that, apart from w_* , w_1 also satisfies (3.15). Let $(x_1, u_1, \eta_1) := \mathcal{F}(\xi, w_1)$. We find from (3.17) that

$$(3.19) \quad \begin{aligned} &\|z_{u_*, w_*, \xi}(T)\|^2 - \|w_*(T)\|^2 - 2\eta_*^T(T)x_*(T+1) + 2\eta_*^T(T-1)x_*(T) \\ &= \|w_*(T)\|^2 - \|z_{u_*, w_*, \xi}(T)\|^2 \end{aligned}$$

We also find that

$$(3.20) \quad \begin{aligned} & \|z_{u_1, w_1, \xi}(T)\|^2 - \|w_1(T)\|^2 - 2\eta_*^\top(T)x_1(T+1) + 2\eta_*^\top(T-1)x_1(T) \\ &= \|z_{u_1, w_1, \xi}(T)\|^2 - \|w_1(T)\|^2 + 2w_*^\top(T)w_1(T) - 2z_{u_*, w_*, \xi}^\top(T)z_{u_1, w_1, \xi}(T). \end{aligned}$$

Summing (3.19) and (3.20) from 0 to ∞ and subtracting from each other gives us

$$(3.21) \quad \mathcal{C}(\xi, w_*) = \mathcal{C}(\xi, w_1) - \|w_* - w_1\|_C^2.$$

Since w_* was maximizing, we find that $\|w_* - w_1\|_C = 0$ and hence $w_* = w_1$. \square

Next, we show that $\mathcal{C}^*(\xi) = \xi^\top P \xi$ for some matrix P . To do that we first show that u_* , η_* , and w_* are linear functions of x_* in the result below.

LEMMA 3.7. *There exist constant matrices K_1 , K_2 , and K_3 such that*

$$(3.22) \quad u_* = K_1 x_*,$$

$$(3.23) \quad \eta_* = K_2 x_*,$$

$$(3.24) \quad w_* = K_3 x_*.$$

Proof. First we look at time 0. By Lemma 3.6 it is easily seen that $\mathcal{H} : \xi \rightarrow w_*$ is linear. Hence the mapping from ξ to $w_*(0)$ is also linear. This implies the existence of a matrix K_3 such that $w_*(0) = K_3 \xi$. From (3.10) and Lemma 3.4, it is easily seen that u_* and η_* are linear functions of ξ and w_* . This implies, since w_* is a linear function of ξ , that $u_*(0)$ and $\eta_*(0)$ are linear functions of ξ , and hence there exist K_1 and K_2 such that $u_*(0) = K_1 \xi$ and $\eta_*(0) = K_2 \xi$.

We now look at time t . The sup-inf problem starting at time t with initial value $x(t)$ can now be solved. Due to time invariance, we see that w_* restricted to $[t, \infty)$ satisfies (3.15), and hence for this problem the optimal x and η are x_* and η_* . However, since t is the initial time for this optimization problem, which is exactly equal to the original optimization problem, we find (3.22)–(3.24) at time t with the same matrices K_1 , K_2 , and K_3 as at time 0. Since t was arbitrary this completes the proof. \square

LEMMA 3.8. *There exists a matrix P such that $\sigma^{-1}\eta_* = -Px_*$. Moreover, for this P we find that*

$$(3.25) \quad \mathcal{C}^*(\xi) = \xi^\top P \xi.$$

Proof. We have

$$\begin{aligned} \sigma^{-1}\eta_* &= [A^\top \eta_* - C_2^\top C_2 x_* - C_2^\top D_{22} w_*] \\ &= (A^\top K_2 - C_2^\top C_2 - C_2^\top D_{22} K_3) x_*. \end{aligned}$$

We define $P := -(A^\top K_2 - C_2^\top C_2 - C_2^\top D_{22} K_3)$ using the matrices defined in lemma 3.7. We prove that this P satisfies (3.25). We can derive the following equation:

$$\begin{aligned} & \|z_{u_*, w_*, \xi}(T)\|^2 - \|w_*(T)\|^2 - 2\eta_*^\top(T)x_*(T+1) + 2\eta_*^\top(T-1)x_*(T) \\ &= \|w_*(T)\|^2 - \|z_{u_*, w_*, \xi}(T)\|^2 \end{aligned}$$

We sum this equation from zero to infinity. Since $\lim_{T \rightarrow \infty} \eta_*(T) = 0$ and $\lim_{T \rightarrow \infty} x_*(T) = 0$, we find that

$$\mathcal{C}(\xi, w_*) + 2\eta_*^\top(-1)x_*(0) = -\mathcal{C}(\xi, w_*).$$

Since $\mathcal{C}(\xi, w_*) = \mathcal{C}^*(\xi)$ and $\eta_*(-1) = -P\xi$, we find (3.25). \square

Next, we show that this matrix P satisfies conditions (a)–(c) of theorem 2.1. We first show part (a). Since we do not yet know if P is symmetric, we must be careful. This essential step in our derivation is new compared to the method for the continuous time as used in [24].

LEMMA 3.9. *Let P be given by Lemma 3.8. The matrices V and R as defined in part (ii) of theorem 2.1, condition (a), satisfy $V + V^T > 0$, $R + R^T > 0$.*

Proof. By Lemmas 3.5 and 3.8, we know that $(P + P^T)/2 \geq L$, and therefore we find that $(V + V^T)/2 \geq D_{21}^T D_{21} + B^T L B$. The latter matrix is positive definite, and hence $(V + V^T)/2$ is positive definite, i.e., $V + V^T > 0$.

We now look at the following “sup-inf-sup-inf” problem for initial condition 0:

$$(3.26) \quad \mathcal{J}(0) := \sup_{w(0)} \inf_{u(0)} \sup_{w^+} \inf_{u^+} \|z_{u,w}\|_2^2 - \|w\|_2^2,$$

where $w^+ := w|_{[1,\infty)}$ and $u^+ := u|_{[1,\infty)}$. We will always add the constraint that u^+ is such that the resulting state x is in ℓ_2 .

We know that there exists a causal operator f satisfying Condition 3.2, and hence this function makes the ℓ_2 -induced operator norm strictly less than 1 under the constraint $x \in \ell_2^n$. In (3.26) we set $u = f(w)$. This is possible since by causality we know that $u(0)$ only depends on $w(0)$ and u^+ depends on the whole function w . Thus we get

$$(3.27) \quad \begin{aligned} \mathcal{J}(0) &= \sup_{w(0)} \inf_{u(0)} \sup_{w^+} \inf_{u^+} \|z_{u,w}\|_2^2 - \|w\|_2^2 \\ &\leq \sup_w \|z_{f(w),w}\|_2^2 - \|w\|_2^2 \\ &\leq 0. \end{aligned}$$

Since, by Lemma 3.8, we have

$$(3.28) \quad \sup_{w^+} \inf_{u^+} \|z_{u^+,w^+,x(1)}\|_2^2 - \|w^+\|_2^2 = x^T(1) P x(1),$$

we can reduce (3.26) to the following “sup-inf” problem:

$$\sup_{w(0)} \inf_{u(0)} \begin{pmatrix} u(0) \\ w(0) \end{pmatrix}^T \begin{pmatrix} V & B^T P E + D_{21}^T D_{22} \\ E^T P B + D_{22}^T D_{21} & E^T P E + D_{22}^T D_{22} - I \end{pmatrix} \begin{pmatrix} u(0) \\ w(0) \end{pmatrix}.$$

When we define

$$\tilde{u}(0) = u(0) - (E^T P B + D_{22}^T D_{21}) V^{-1} w(0),$$

we get

$$(3.29) \quad \mathcal{J}(0) = \sup_{w(0)} \inf_{\tilde{u}(0)} \begin{pmatrix} \tilde{u}(0) \\ w(0) \end{pmatrix}^T \begin{pmatrix} V & 0 \\ 0 & -R \end{pmatrix} \begin{pmatrix} \tilde{u}(0) \\ w(0) \end{pmatrix}.$$

Since, by (3.27), $\mathcal{J}(0)$ is finite we immediately find that a necessary condition is $R + R^T \geq 0$.

Assume that $R + R^T$ is not invertible. Then there exists a $v \neq 0$ such that $v^T R v = 0$. Let $w^+(u(0))$ be the ℓ_2 -function that attains the optimum in the optimization (3.28) with initial state $x(1) = B u(0) + E v$. We define the function w by

$$(3.30) \quad [w(u(0))](t) := \begin{cases} v & \text{if } t = 0, \\ w^+(u(0))(t) & \text{otherwise.} \end{cases}$$

Assume that δ and f are such that Condition 3.2 is satisfied. Define u by

$$(3.31) \quad u = f[w(u(0))].$$

Since the map from u to w defined by (3.30) is strictly causal and since f is causal, u is uniquely defined by (3.31). To prove this, note that $u(0)$ only depends on $w(u(0))(0) = v$, and hence w^+ as a function of $u(0)$ is uniquely defined, which in turn yields u . Denote u and w obtained in this way by u_1 and w_1 . By (3.27) and (3.29), we find that, for this particular choice of w_1 and u_1 , we have

$$(3.32) \quad \|z_{u_1, w_1}\|_2^2 - \|w_1\|_2^2 \geq 0.$$

On the other hand, using Condition 3.2 we find that

$$\|z_{u_1, w_1}\|_2^2 - \|w_1\|_2^2 < (\delta^2 - 1) \|w_1\|_2^2 < 0$$

since $w_1(0) = v \neq 0$. Therefore we have a contradiction, and hence our assumption that $R + R^T$ is not invertible was incorrect. Together with $R + R^T \geq 0$, this yields $R + R^T > 0$. \square

LEMMA 3.10. *Assume that (A, B, C_2, D_{21}) has no invariant zeros on the unit circle and is left-invertible. Moreover, assume that $D_{21}^T[C_2 \ D_{22}] = 0$. If the condition in part (i) of Theorem 2.1 is satisfied, then there exists a symmetric matrix $P \geq 0$ satisfying (a)–(c) of part (ii) of Theorem 2.1.*

Proof. We define the matrices

$$\begin{aligned} M &:= D_{21}^T D_{21} + B^T L B > 0, \\ Z &:= I - D_{22}^T D_{22} - E^T X_1 L E. \end{aligned}$$

We know that $-(R + R^T)/2$ is the Schur complement of $(V + V^T)/2$ in $G((P + P^T)/2)$. By Lemma 3.9, we know that $R + R^T > 0$ and $V + V^T > 0$. Therefore $G((P + P^T)/2)$ has m eigenvalues on the positive real axis and l eigenvalues on the negative real axis. We know that $G((P + P^T)/2) - G(L) \geq 0$ since $(P + P^T)/2 \geq L$. An easy consequence of the theorem of Courant–Fischer (see [2]) then tells us that $G(L)$ has at least l eigenvalues on the negative real axis. Since $-Z$ is the Schur complement of $M > 0$ in $G(L)$, this implies that $Z > 0$.

By Lemma 3.8, we have $\eta_* = -\sigma P x_*$. By combining Lemmas 3.4 and 3.6 and rewriting the equations, we find that u_* and w_* satisfy the following equations:

$$\begin{aligned} w_* &= Z^{-1} \{E^T X_1 (P - L) \sigma x_* + (D_{22}^T C_2 + E^T X_1 L A) x_*\}, \\ u_* &= -M^{-1} B^T \{(P - L) \sigma x_* + L A x_* + L E w_*\}. \end{aligned}$$

Thus we get

$$(3.33) \quad \begin{aligned} &\{I + [B M^{-1} B^T - X_1^T E Z^{-1} E^T X_1] (P - L)\} x_*(k+1) \\ &= X_1^T \{A + E Z^{-1} E^T X_1 L A + E Z^{-1} D_{22}^T C_2\} x_*(k) \end{aligned}$$

Since, by Lemma 3.9, R as defined in Theorem 2.1 is invertible, it can be shown that the matrix on the left is invertible, and hence (3.33) uniquely defines $x_*(k+1)$ as a function of $x_*(k)$. It follows that (3.33) can be rewritten in the form $\sigma x_* = A_{cl} x_*$, with A_{cl} as defined by (2.7). Since $x_* \in \ell_2^n$ for every initial state ξ , we know that A_{cl} is asymptotically stable. Next, we show that P satisfies the discrete algebraic

Riccati equation (2.6). From the backward difference equation in (3.10) combined with Lemma 3.8 and the formula given above for w_* , we find that

$$P = A^T P A_{cl} + C_2^T C_2 + C_2^T D_{22} Z^{-1} \{E^T X_1 (P - L) A_{cl} + D_{22}^T C_2 + E^T X_1 L A\}$$

By some extensive calculations this equation turns out to be equivalent to the discrete algebraic Riccati equation (2.6). Next we show that P is symmetric. Note that both P and P^T satisfy the discrete algebraic Riccati equation. Using this we find that $(P - P^T) = A_{cl}^T (P - P^T) A_{cl}$. Since A_{cl} is asymptotically stable, this implies that $P = P^T$. P can be shown to be positive semidefinite by combining Lemma 3.5 and (3.25). It remains to be shown that P satisfies (2.4). Since P is symmetric, we know that V and R are symmetric. Equation (2.4) is then an immediate consequence of lemma 3.9. \square

We extend this result in the following corollary to systems that do not satisfy (3.1).

COROLLARY 3.11. *Assume that (A, B, C_2, D_{21}) has no invariant zeros on the unit circle and is left invertible. If part (i) of Theorem 2.1 is satisfied, then there exists a symmetric matrix $P \geq 0$ satisfying (a)–(c) of part (ii) of Theorem 2.1.*

Proof. We first apply a preliminary feedback $u = \tilde{F}_1 x + \tilde{F}_2 w + v$ such that $D_{21}^T (C_2 + D_{21} \tilde{F}_1) = 0$ and $D_{21}^T (D_{22} + D_{21} \tilde{F}_2) = 0$. Denote the new A , C_2 , D_{22} , and E by \tilde{A} , \tilde{C}_2 , \tilde{D}_{22} , and \tilde{E} . For this new system, Condition 3.2 is satisfied. We also know that by applying a preliminary state feedback, the invariant zeros of a system do not change. Therefore the new subsystem $(\tilde{A}, B, \tilde{C}_2, D_{21})$ does not have invariant zeros on the imaginary axis. Hence, since for this new system $D_{21}^T [\tilde{C}_2 \quad \tilde{D}_{22}] = 0$, we find conditions in terms of the new parameters by applying Lemma 3.10. Rewriting in terms of the original parameters gives the desired conditions (a)–(c) as given in part (ii) of Theorem 2.1. \square

4. A first system transformation. To proceed with the proof of Theorem 2.1, (i) \Rightarrow (ii), in this section we will transform our original system (2.1) into a new system. The problem of finding an internally stabilizing feedback that makes the H_∞ norm of the closed loop system less than 1 for the original system is equivalent to the problem of finding an internally stabilizing feedback that makes the H_∞ norm of the closed loop system less than 1 for the new transformed system. However, this new system has some very desirable properties, which makes working with it much easier. In particular, for this new system the disturbance decoupling problem with measurement feedback is solvable. We will perform the transformation in two steps.

First we will perform a transformation related to the full-information H_∞ problem, and next a transformation related to the filtering problem.

We assume that we have a positive semidefinite matrix P satisfying conditions (a)–(c) of Theorem 2.1. By the results of the previous section, this matrix exists in the case where part (i) of Theorem 2.1 is satisfied. We define the following system:

$$(4.1) \quad \Sigma_P : \begin{cases} \sigma x_P = A_P x_P + B u_P + E_P w_P, \\ y_P = C_{1,P} x_P + \quad \quad \quad + D_{12,P} w_P, \\ z_P = C_{2,P} x_P + D_{21,P} u_P + D_{22,P} w_P, \end{cases}$$

where the matrices are as defined in the statement of Theorem 2.1. Furthermore, we define the following system:

$$(4.2) \quad \Sigma_U : \begin{cases} \sigma x_U = A_U x_U + B_U u_U + E_U w, \\ y_U = C_{1,U} x_U + \quad \quad \quad + D_{12,U} w, \\ z_U = C_{2,U} x_U + D_{21,U} u_U + D_{22,U} w, \end{cases}$$

where

$$\begin{aligned}
A_U &:= A - BV^{-1}(B^T P A + D_{21}^T C_2), \\
B_U &:= BV^{-1/2}, \\
E_U &:= E - BV^{-1}(B^T P E + D_{21}^T D_{22}), \\
C_{1,U} &:= -R^{-1/2} H, \\
C_{2,U} &:= C_2 - D_{21} V^{-1}(B^T P A + D_{21}^T C_2), \\
D_{12,U} &:= R^{1/2}, \\
D_{21,U} &:= D_{21} V^{-1/2}, \\
D_{22,U} &:= D_{22} - D_{21} V^{-1}(B^T P E + D_{21}^T D_{22}),
\end{aligned}$$

and V , R , and H are as defined in Theorem 2.1. We will show that Σ_U has a very nice property. To do this, we will first give a definition and some results we will need in the sequel. A system is called *inner* if the system is internally stable, square (i.e., the number of inputs is equal to the number of outputs) and the transfer matrix of the system, denoted by G , satisfies

$$(4.3) \quad G(z)G^T(z^{-1}) = I$$

We will now formulate a generalization of [12, Lemma 5] to the case where $G(z)$ may have poles in zero. The proof is slightly more complicated than the one given in [12], since if G has a pole in zero then $G^T(z^{-1})$ is no longer proper. Nevertheless, a proof can be given by simply writing out (4.3).

LEMMA 4.1. *Assume that we have a square system*

$$(4.4) \quad \Sigma_{st} : \begin{cases} \sigma x = Ax + Bu, \\ z = Cx + Du. \end{cases}$$

Assume that A is asymptotically stable. The system Σ_{st} is inner if there exists a matrix X satisfying

1. $X = A^T X A + C^T C$,
2. $D^T C + B^T X A = 0$,
3. $D^T D + B^T X B = I$.

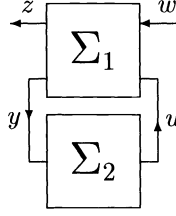
Remarks.

(i) If (A, B) is controllable, the reverse of the above implication is also true. However, in general, the reverse does not hold. A simple counterexample is given by $\Sigma_{st} := (0.5, 0, 1, 1)$, which is inner but for which (ii) does not hold for any choice of X .

(ii) Note that if a matrix X satisfies part (1) of Lemma 4.1, then it is equal to the observability gramian of (C, A) . We know, for instance, that $X > 0$ if and only if (C, A) is observable. In general, we only have $X \geq 0$.

We have the following important property of inner systems (see [17], [22]).

LEMMA 4.2. *Suppose that we have the following interconnection of two systems Σ_1 and Σ_2 , both described by some state-space representation:*



(4.5)

Assume that Σ_1 is inner. Denote its transfer matrix from (w, u) to (z, y) by L . Moreover, assume that if we decompose L compatible with the sizes of w, u, z , and y :

$$(4.6) \quad L \begin{pmatrix} w \\ u \end{pmatrix} =: \begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix} \begin{pmatrix} w \\ u \end{pmatrix} = \begin{pmatrix} z \\ y \end{pmatrix},$$

we have $L_{21}^{-1} \in H_\infty$, and L_{22} is strictly proper. Then the following two statements are equivalent:

(i) The closed loop system (4.5) is internally stable and its closed loop transfer matrix has H_∞ norm less than 1.

(ii) The system Σ_2 is internally stable and its transfer matrix has H_∞ norm less than 1.

LEMMA 4.3. The system Σ_U as defined by (4.2) is inner. Denote the transfer matrix of Σ_U by U . We decompose U compatible with the sizes of w, u_U, z_U , and y_U :

$$U \begin{pmatrix} w \\ u_U \end{pmatrix} =: \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix} \begin{pmatrix} w \\ u_U \end{pmatrix} = \begin{pmatrix} z_U \\ y_U \end{pmatrix}.$$

Then U_{21} is invertible and its inverse is in H_∞ . Moreover, U_{22} is strictly proper.

Proof. It can be easily checked that P as defined by Theorem 2.1 (a)–(c) satisfies conditions (i)–(iii) of Lemma 4.1. Part (i) of Lemma 4.1 turns out to be equal to the discrete algebraic Riccati equation (2.6). Parts (ii) and (iii) follow by simply writing out the equations in terms of the original system parameters of system (2.1).

Next, we show that A_U is asymptotically stable. We know $P \geq 0$ and

$$(4.7) \quad P = A_U^T P A_U + \begin{pmatrix} C_{1,U}^T & C_{2,U}^T \end{pmatrix} \begin{pmatrix} C_{1,U} \\ C_{2,U} \end{pmatrix}.$$

It can be easily checked that $x \neq 0$, $A_U x = \lambda x$, $C_{1,U} x = 0$, and $C_{2,U} x = 0$ imply that $A_{cl,P} x = \lambda x$, where $A_{cl,P}$ is defined by (2.7). Since $A_{cl,P}$ is asymptotically stable, we have $\text{Re } \lambda < 0$. Hence the realization (4.2) is detectable. By standard Lyapunov theory, the existence of a positive semidefinite solution of (4.7), together with detectability, guarantees asymptotic stability of A_U .

We can immediately write down the following realization for U_{21}^{-1} :

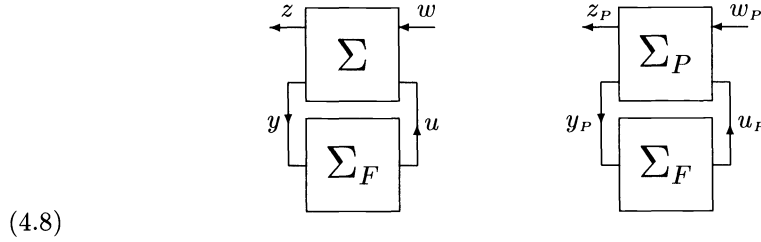
$$\Sigma_{U_{21}^{-1}} : \begin{cases} \sigma x_U = (A_U - E_U D_{12,U}^{-1} C_{1,U}) x_U + E_U D_{12,U}^{-1} w, \\ y_U = -D_{12,U}^{-1} C_{1,U} x_U + D_{12,U}^{-1} w. \end{cases}$$

Since $A_{cl,P} = A_U - E_U D_{12,U}^{-1} C_{1,U}$ we know that U_{21}^{-1} is an H_∞ function.

Finally, the claim that U_{22} is strictly proper is trivial to check. This completes the proof. \square

We will now formulate our key lemma, below.

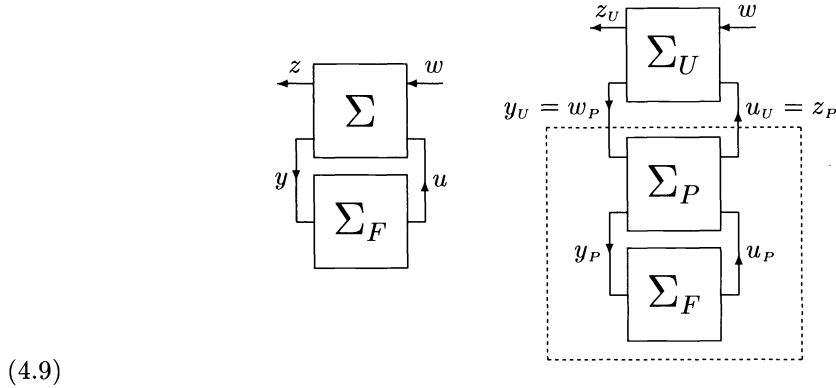
LEMMA 4.4. *Let P satisfy Theorem 2.1 part (ii), (a)–(c). Moreover, let Σ_F be an arbitrary linear time-invariant finite-dimensional compensator in the form (2.2). Consider the following two systems, where the system on the left is the interconnection of (2.1) and (2.2), and the system on the right is the interconnection of (4.1) and (2.2):*



Then the following statements are equivalent

- (i) The system on the left is internally stable and its transfer matrix from w to z has H_∞ norm less than 1.
- (ii) The system on the right is internally stable and its transfer matrix from w_P to z_P has H_∞ norm less than 1.

Proof. We investigate the following systems:



The system on the left is the same as the system on the left in (4.8), and the system on the right is described by system (4.2) interconnected with the system on the right in (4.8). A realization for the system on the right is given by

$$\sigma \begin{pmatrix} x_U - x_{1,P} \\ x_P \\ p \end{pmatrix} = \begin{pmatrix} A_{cl,P} & 0 & 0 \\ * & A + BNC_1 & BM \\ * & LC_1 & K \end{pmatrix} \begin{pmatrix} x_U - x_{1,P} \\ x_P \\ p \end{pmatrix} + \begin{pmatrix} 0 \\ E + BND_{12} \\ LD_{12} \end{pmatrix} w,$$

$$z_U = \begin{pmatrix} * & C_2 + D_{21}NC_1 & D_{21}M \end{pmatrix} \begin{pmatrix} x_U - x_{1,P} \\ x_P \\ p \end{pmatrix} + (D_{22} + D_{21}ND_{12}) w$$

where $A_{cl,P}$ is defined by (2.7). The $*$'s denote matrices that are unimportant for this argument. The system on the right is internally stable if and only if the system

described by the above set of equations is internally stable. If we also derive the system equations for the system on the left in (4.9), we immediately see that, since $A_{cl,P}$ is asymptotically stable, the system on the left is internally stable if and only if the system on the right is internally stable. Moreover, if we take zero initial conditions and both systems have the same input w then we have $z = z_U$; i.e., the input-output behaviour of both systems are equivalent. Hence the system on the left has H_∞ norm less than 1 if and only if the system on the right has H_∞ norm less than 1.

By Lemma 4.3, we may apply Lemma 4.2 to the system on the right in (4.9), and hence we find that the closed loop system is internally stable and has H_∞ norm less than 1 if and only if the dashed system is internally stable and has H_∞ norm less than 1.

Since the dashed system is exactly the system on the right in (4.8) and the system on the left in (4.9) is exactly equal to the system on the left in (4.8), we have completed the proof. \square

Using the previous lemma, we know that we only have to investigate the system Σ_P . This new system has some very nice properties, which we will use. First, we will look at the Riccati equation for the system Σ_P . It can be checked immediately that $X = 0$ satisfies (a)–(c) of Theorem 2.1 for the system Σ_P .

We now dualize Σ_P . We know that (A, E, C_1, D_{12}) is right-invertible and has no invariant zeros on the unit circle. It can be easily checked that this implies that $(A_P, E, C_{1,P}, D_{12})$ is right-invertible and has no invariant zeros on the unit circle. Hence for the dual of Σ_P we know that $(A_P^T, C_{1,P}^T, E^T, D_{21}^T)$ is left-invertible and has no invariant zeros on the unit circle. If there exists an internally stabilizing feedback for the system Σ , which makes the H_∞ norm of the closed loop system less than 1, then the same feedback is internally stabilizing and makes the H_∞ norm of the closed loop system less than 1 for the system Σ_P . If we dualize this feedback and apply it to the dual of Σ_P , then it is again internally stabilizing and again it makes the H_∞ norm of the closed loop system less than 1. We can now apply the dual version of Corollary 3.11, which exactly guarantees the existence of a matrix Y satisfying conditions (d)–(f) of Theorem 2.1. Thus we derived the following lemma, which gives the necessity part of Theorem 2.1.

LEMMA 4.5. *Let system (2.1) be given with zero initial state. Assume that (A, B, C_2, D_{21}) has no invariant zeros on the unit circle and is left-invertible. Moreover, assume that (A, E, C_1, D_{12}) has no invariant zeros on the unit circle and is right invertible. If part (i) of theorem 2.1 is satisfied, then there exist matrices P and Y satisfying (a)–(f) of part (ii) of Theorem 2.1.*

This completes the proof (i) \Rightarrow (ii). In the next section we will prove the reverse implication. Moreover, in the case where the desired feedback exists, we will derive an explicit formula for one choice for Σ_F that satisfies all requirements.

5. The transformation into a disturbance decoupling problem with measurement feedback. In this section we will assume that there exist matrices P and Y satisfying part (ii) of Theorem 2.1 for system (2.1). We will transform our original system Σ into another system $\Sigma_{P,Y}$. We will show that a compensator is internally stabilizing and makes the H_∞ norm of the closed loop system less than 1 for the system Σ if and only if the same compensator is internally stabilizing and makes the H_∞ norm of the closed loop system less than 1 for our transformed system $\Sigma_{P,Y}$. After that, we will show that $\Sigma_{P,Y}$ has the following very special property (see [19]):

There exists an internally stabilizing compensator that makes the closed loop transfer matrix equal to zero; i.e., w does not have any effect on the output of the system z . This property of $\Sigma_{P,Y}$ has a special name: “the disturbance decoupling problem with measurement feedback and internal stability (DDPMS).”

We first define $\Sigma_{P,Y}$. We start by transforming Σ into Σ_P . Then we apply the dual transformation on Σ_P to obtain $\Sigma_{P,Y}$:

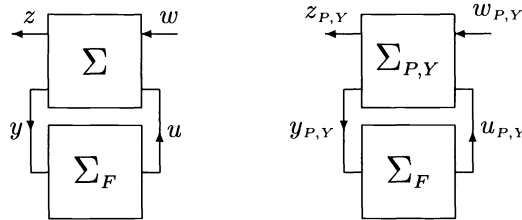
$$(5.1) \quad \Sigma_{P,Y} : \begin{cases} \sigma x_{P,Y} = A_{P,Y} x_{P,Y} + B_{P,Y} u_{P,Y} + E_{P,Y} w_{P,Y}, \\ y_{P,Y} = C_{1,P} x_{P,Y} + D_{12,P} w_{P,Y}, \\ z_{P,Y} = C_{2,P} x_{P,Y} + D_{21,P} u_{P,Y} + D_{22,P} w_{P,Y}, \end{cases}$$

where

$$\begin{aligned} \tilde{H} &:= A_P Y C_{2,P}^T + E_P D_{22,P}^T - (A_P Y C_{1,P}^T + E_P D_{12,P}^T) W^{-1} \times \\ &\quad (C_{1,P} Y C_{2,P}^T + D_{12,P} D_{22,P}^T), \\ A_{P,Y} &:= A_P + \tilde{H} S^{-1} C_{2,P}, \\ C_{2,P,Y} &:= S^{-1/2} C_{2,P}, \\ B_{P,Y} &:= B + \tilde{H} S^{-1} D_{21,P}, \\ E_{P,Y} &:= (A_P Y C_{1,P}^T + E_P D_{12,P}^T) W^{-1/2} + \tilde{H} S^{-1} (C_{2,P} Y C_{1,P}^T + D_{22,P} D_{12,P}^T) W^{-1/2}, \\ D_{12,P,Y} &:= W^{1/2}, \\ D_{21,P,Y} &:= S^{-1/2} D_{21,P}, \\ D_{22,P,Y} &:= S^{-1/2} (C_{2,P} Y C_{1,P}^T + D_{22,P} D_{12,P}^T) W^{-1/2}. \end{aligned}$$

When we first apply Lemma 4.4 on the transformation from Σ to Σ_P and then the dual of Lemma 4.4 on the transformation from Σ_P to $\Sigma_{P,Y}$, we find the following result.

LEMMA 5.1. *Let P satisfy Theorem 2.1, part (ii) (a)–(c). Moreover, let an arbitrary linear time-invariant finite-dimensional compensator Σ_F be given, described by (2.2). Consider the following two systems, where the system on the left is the interconnection of (2.1) and (2.2), and the system on the right is the interconnection of (5.1) and (2.2):*



Then the following statements are equivalent:

(i) *The system on the left is internally stable and its transfer matrix from w to z has H_∞ norm less than 1.*

(ii) *The system on the right is internally stable and its transfer matrix from $w_{P,Y}$ to $z_{P,Y}$ has H_∞ norm less than 1.*

It remains to be shown that for $\Sigma_{P,Y}$ the DDPMS is solvable.

LEMMA 5.2. Let Σ_F be given by

$$(5.2) \quad \Sigma_F : \begin{cases} \sigma p &= K_{P,Y} p + L_{P,Y} y_{P,Y}, \\ u_{P,Y} &= M_{P,Y} p + N_{P,Y} y_{P,Y}, \end{cases}$$

where

$$\begin{aligned} N_{P,Y} &:= -D_{21,P,Y}^{-1} D_{22,P,Y} D_{12,P,Y}^{-1}, \\ M_{P,Y} &:= -(D_{21,P,Y}^{-1} C_{2,P,Y} + N_{P,Y} C_{1,P}), \\ L_{P,Y} &:= B_{P,Y} N_{P,Y} + E_{P,Y} D_{12,P,Y}^{-1}, \\ K_{P,Y} &:= A_{P,Y} + B_{P,Y} M_{P,Y} - E_{P,Y} D_{12,P,Y}^{-1} C_{1,P}. \end{aligned}$$

The interconnection of Σ_F and $\Sigma_{P,Y}$ is internally stable, and the closed loop transfer matrix from $w_{P,Y}$ to $z_{P,Y}$ is zero.

Proof. We can write out the formulas for a state-space representation of the interconnection of $\Sigma_{P,Y}$ and Σ_F . We then apply the following basis transformation:

$$\begin{pmatrix} x_{P,Y} - p \\ p \end{pmatrix} = \begin{pmatrix} I & -I \\ 0 & I \end{pmatrix} \begin{pmatrix} x_{P,Y} \\ p \end{pmatrix}.$$

After this transformation we immediately see that the closed loop transfer matrix from $w_{P,Y}$ to $z_{P,Y}$ is zero. Moreover, the system matrix (2.3) after this transformation is given by:

$$\begin{pmatrix} A_{cl,P,Y} & 0 \\ L_{P,Y} C_{1,P} & A_{cl,P} \end{pmatrix}$$

Since $A_{cl,P,Y}$ and $A_{cl,P}$ are asymptotically stable matrices, this implies that, indeed, Σ_F is internally stabilizing. \square

This controller is the same as the one described in the statement of Theorem 2.1. We know that Σ_F is internally stabilizing, and the resulting closed loop system has H_∞ norm less than 1 for the system $\Sigma_{P,Y}$. Hence, by applying Lemma 5.1, we find that Σ_F satisfies part (i) of Theorem 2.1. This completes the proof of (ii) \Rightarrow (i) of Theorem 2.1. We have already shown the reverse implication and hence the proof of Theorem 2.1 is complete.

6. Conclusions. In this paper we have solved the discrete time H_∞ problem with measurement feedback. It is shown that the techniques for the continuous time case can be applied to the discrete time case. Unfortunately, the formulas are much more complex, but it is still possible to give an explicit formula for one controller satisfying all requirements. It would, however, be interesting to generalize this work and find a characterization of all controllers satisfying the requirements. Another interesting open problem is to derive recursive formulas to calculate the solutions to these algebraic Riccati equations. It would also be interesting to find two dual Riccati equations and a coupling condition, as in [9]. Nevertheless, the results presented in this paper show that it is very possible to solve discrete time H_∞ problems directly, instead of transforming them to continuous time. The assumption of left-invertibility is not very restrictive. It implies that there are several inputs that have the same effect on the output and this nonuniqueness can be factored out (see [18] for a continuous time treatment). The assumption of right-invertibility can be removed by dualizing this reasoning. However, at this moment it is unclear as to how to remove the assumptions concerning zeros on the unit circle.

REFERENCES

- [1] T. BAŞAR, *A dynamic games approach to controller design: disturbance rejection in discrete time*, Proc. CDC, Tampa, 1989, pp. 407–414.
- [2] R. BELLMAN, *Introduction to matrix analysis*, 2nd ed., McGraw-Hill, New York, 1970.
- [3] T. CHEN AND B. A. FRANCIS, \mathcal{H}_2 -optimal sampled-data control, IEEE Trans. Aut. Contr., 36 (1991), pp. 387–397.
- [4] ———, *On the \mathcal{L}_2 -induced norm of a sampled-data system*, Syst. & Contr. Letters Vol. 15, 1990, pp. 211–220.
- [5] J. C. DOYLE, *Lecture notes in advances in multivariable control*, ONR/Honeywell Workshop, Minneapolis, 1984.
- [6] J. DOYLE, K. GLOVER, P. P. KHARGONEKAR, AND B. A. FRANCIS, *State space solutions to standard H_2 and H_∞ control problems*, IEEE Trans. Aut. Contr., 34, 1989, pp. 831–847.
- [7] B. A. FRANCIS, *A course in H_∞ control theory*, Lecture notes in control and information sciences, Vol. 88, Springer-Verlag, Berlin, 1987.
- [8] Y. GENIN, P. VAN DOOREN, T. KAILATH, J. M. DELOSME, AND M. MORF, *On Σ -lossless transfer functions and related questions*, Lin. Alg. Appl., 50, 1983, pp. 251–275.
- [9] K. GLOVER AND J. DOYLE, *State-space formulae for all stabilizing controllers that satisfy an H_∞ norm bound and relations to risk sensitivity*, Syst. & Contr. Letters, 11, 1988, pp. 167–172.
- [10] K. GLOVER, *All optimal Hankel-norm approximations of linear multivariable systems and their L^∞ -error bounds*, Int. J. Contr., 39, 1984, pp. 1115–1193.
- [11] M. J. GRIMBLE, *Optimal robustness and the relationship to LQG design problems*, Int. J. Contr., 43, 1986, pp. 351–372.
- [12] D. W. GU, M. C. TSAI, AND I. POSTELTHWAITE, *State space formulae for discrete time H_∞ optimization*, Int. J. Contr., 49, 1989, pp. 1683–1723.
- [13] P. P. KHARGONEKAR, I. R. PETERSEN, AND M. A. ROTE, *H_∞ optimal control with state feedback*, IEEE Trans. Aut. Contr., 33, 1988, pp. 786–788.
- [14] H. KWAKERNAAK, *A polynomial approach to minimax frequency domain optimization of multivariable feedback systems*, Int. J. Contr., 41, 1986, pp. 117–156.
- [15] E. B. LEE AND L. MARKUS, *Foundations of optimal control theory*, Wiley, New York, 1967.
- [16] D. J. N. LIMBEER, M. GREEN, AND D. WALKER, *Discrete time H_∞ control*, Proc. CDC, Tampa, 1989, pp. 392–396.
- [17] R. M. REDHEFFER, *On a certain linear fractional transformation*, J. Math. and Physics, 39, 1960, pp. 269–286.
- [18] C. SCHERER, *H_∞ control by state feedback and fast algorithms for the computation of optimal H_∞ norms*, IEEE Trans. Aut. Contr., 35, 1990, pp. 1090–1099.
- [19] J. M. SCHUMACHER, *Dynamic feedback in finite and infinite dimensional linear systems*, Math. Centre Tracts 143, Amsterdam, 1981.
- [20] L. M. SILVERMAN, *Discrete Riccati equations: alternative algorithms, asymptotic properties and system theory interpretation*, in Control and dynamic systems, Academic, New York, 12, 1976, pp. 313–386.
- [21] A. A. STOORVOGEL AND H. L. TRENTelman, *The quadratic matrix inequality in singular H_∞ control with state feedback*, SIAM J. Contr. & Opt., 28 (1990), pp. 1190–1208.
- [22] A. A. STOORVOGEL, *The singular H_∞ control problem with dynamic measurement feedback*, SIAM J. Control Optim. 29 (1991), pp. 160–184.
- [23] ———, *The discrete time H_∞ control problem: the full information case*, COSOR memorandum 89-25, Eindhoven University of Technology, October 1989.
- [24] G. TADMOR, *Worst case design in the time domain: the maximum principle and the standard H_∞ problem*, Math. Contr. Sign. & Syst., 3 (1990), pp. 301–324.
- [25] H. L. TRENTelman AND A. A. STOORVOGEL, *Completion of the squares in the finite horizon H_∞ control problem by measurement feedback*, Proc conf. on New trends in system theory, Genova, 1990, to appear.
- [26] P. WHITTLE, *Risk-sensitive linear/quadratic/Gaussian control*, Adv. Appl. Prob., 13 (1981), pp. 764–777.
- [27] I. YAESH AND U. SHAKED, *Minimum H_∞ -norm regulation of linear discrete-time systems and its relation to linear quadratic discrete games*, IEEE Trans. Aut. Contr., 35 (1990), pp. 1061–1064.
- [28] G. ZAMES, *Feedback and optimal sensitivity: model reference transformations, multiplicative seminorms, and approximate inverses*, IEEE Trans. Aut. Contr., 26 (1981), pp. 301–320.

NEW RESULTS IN POLE ASSIGNMENT BY REAL OUTPUT FEEDBACK*

JOACHIM ROSENTHAL†

Abstract. This paper considers the problem of tuning natural frequencies of a linear system by a memoryless controller. Using algebro-geometric methods it is shown how it is possible to improve current sufficiency conditions.

The main result is an exact combinatorial characterization of the nilpotency index of the mod 2 cohomology ring of the real Grassmannian. Using this characterization, new sufficiency results for generic pole assignment for the linear system with m -inputs, p -outputs, and McMillan degree n are given. Among other results it is shown that

$$2.25 \cdot \max(m, p) + \min(m, p) - 3 \geq n$$

is a sufficient condition for generic real pole placement, provided $\min(m, p) \geq 4$.

Key words. output feedback, pole placement, intersection theory, symmetric functions.

AMS(MOS) subject classifications. 93, 55M30, 05A

1. Introduction. Consider a linear time invariant system Σ with m -inputs, p -outputs and McMillan degree n . In the time domain Σ can be modelled by the following system of differential equations:

$$(1.1) \quad \Sigma : \begin{cases} \dot{x} &= Ax + Bu \\ y &= Cx \end{cases} \quad x \in \mathbf{R}^n, y \in \mathbf{R}^p, u \in \mathbf{R}^m.$$

The problem of output pole assignment with a static compensator is the problem of finding a feedback law $u = Fy$ in such a way that the closed loop system

$$(1.2) \quad \Sigma_F : \begin{cases} \dot{x} &= (A + BFC)x \\ y &= Cx \end{cases}$$

is assigned a desired set of eigenvalues. The stability of equilibria or periodic motions of the closed loop system depends on the eigenvalues of the matrix $A + BFC$. In particular the closed loop system is asymptotically stable, if the eigenvalues of $A + BFC$ have negative real parts. In this paper we are interested in under which conditions it is possible to assign a set of real eigenvalues, in particular when it is possible to stabilize a generic system. Because the eigenvalues correspond to the poles of the transfer function under Laplace transform one often refers to this type of problem as the pole placement problem. This question has already been considered by many authors (e.g. [1], [2], [10], [16], [20], [21]), and interesting links to topological questions and Schubert calculus were made. An excellent survey article can be found in [3], where a larger bibliography is also given.

Kimura [10], motivated by the problem of stabilizing and controlling a mechanical system, studied this inverse eigenvalue problem in a systematic way. Typically such systems have m -inputs m -outputs and the dimension of the state is $2m$. More generally, one would hope that $m + p \geq n$ would imply pole assignability, and hence

* Received by the editors May 29, 1990; accepted for publication (in revised form) January 10, 1991.

† Department of Mathematics, University of Notre Dame, Notre Dame, Indiana 46556. This research, which constitutes a part of the author's Ph.D. dissertation at Arizona State University, was done during the author's stay at the Department of Systems Science and Mathematics, Washington University, St. Louis, Missouri 63130.

stabilizability for the generic $p \times m$ n -dimensional system. In 1975, Kimura proved a result, which came “within one degree of freedom” of the desired result.

THEOREM 1.1 (KIMURA [10]). *If (A, B) is controllable and (A, C) is observable and if $m + p - 1 \geq n$, an almost arbitrary set of distinct real or complex conjugate poles is assignable by real gain output feedback.*

In 1978, Willems and Hesselink [21] showed that in the case of $m = p = 2$, at most 3 real poles can be assigned arbitrarily for the generic system, so that Theorem 1.1 also gives a necessary condition for this case.

Quite surprisingly, as shown in this paper, the case studied by Willems and Hesselink is the only nontrivial case ($\min(m, p) \geq 2$) where $m + p \geq n$ is not a sufficient condition. This result will follow from a new combinatorial criterion, which will be formulated in the next section. In fact more will be shown. Using an identification of the mod 2 cohomology ring of the real Grassmannian with a quotient of the space of symmetric functions it will be possible to characterize the maximum number of nontrivial terms in a nonzero product of $H^*(\text{Grass}(p, m + p), \mathbb{Z}_2)$, sometimes called the cup length of this ring, in a combinatorial way. In the next section the combinatorial criterion is formulated and the main results stated.

2. A new combinatorial criterion. Consider a $m \times p$ array A , where m can be seen as the number of inputs and p as the number of outputs. Let $\mu = (\mu_1, \dots, \mu_s)$ be a partition of mp . This means $\mu_1 \geq \mu_2 \geq \dots \geq \mu_s > 0$ and $\sum \mu_i = mp$. Denote with K_μ the number of possibilities to insert μ_i integers i into the array A under the condition that the rows are increasing and the columns are strictly increasing.

DEFINITION 2.1. $c(m, p) := \max\{s \mid K_{(\mu_1, \dots, \mu_s)} \text{ is odd}\}$.

THEOREM 2.2. *The cup length of the mod 2 cohomology ring of the real Grassmannian $\text{Grass}(p, m + p)$ is $c(m, p)$.*

This cup length has an important topological meaning. As was shown by Eilenberg [4], this number gives a lower estimate for the Ljusternik Snirelmann category of a topological space.

In the innovative paper [1], Brockett and Byrnes explained the pole placement problem with static compensators as an intersection problem in some Grassmann variety. Moreover Byrnes [2] showed that the Ljusternik Snirelmann category of the real Grassmannian gives a lower bound for the number of real poles which can be generically assigned. Using Theorem 2.2 therefore, one has immediately the following result.

THEOREM 2.3. *$c(m, p) \geq n$ is a sufficient condition for generic pole placement of a generic, strictly proper linear system Σ_n with m inputs and p outputs and McMillan degree n .*

Clearly not every $m \times p$ system Σ of order n can be pole assigned by output feedback; in particular one needs controllability and observability of the system. The results we present in this paper are stated for a generic system (see, e.g., [21]). Recall that a subset of a variety is called generic if it contains a nonempty Zariski open subset. Before giving the proof of Theorem 2.3, it will be illustrated how it is possible to obtain new sufficiency conditions for generic real pole placement. The following examples were given in [15].

3. Examples and corollaries.

Example 3.1. Two inputs, two outputs or $m = p = 2$. To apply Theorem 2.3, compute K_μ for different partitions of 4:

$K_{(1,1,1,1)} = 2$ (\rightarrow even) given by the two possibilities:

1	2
3	4

1	3
2	4

$K_{(2,1,1)} = 1$ ($\rightarrow odd$) given by the only possibility:

1	1
2	3

Because $K_{(2,1,1)}$ is odd, $c(2,2) = 3$, consistent with the result of Kimura [10] and Willems and Hesselink [21].

Example 3.2. Two inputs, three outputs, or $m = 2$ and $p = 3$. In this case, one immediately computes $K_{(1^6)} = 5$ ($\rightarrow odd$) given by the possibilities:

1	2	3
4	5	6

1	2	4
3	5	6

1	2	5
3	4	6

1	3	4
2	5	6

1	3	5
2	4	6

In other words $c(2,3)=6$ and up to 6 poles can be placed generically. This result is somewhat surprising, although it was already established in the paper of Brockett and Byrnes [1].

Example 3.3. The following table shows $c(m,p)$ for $\max(m,p) \leq 5$.

(3.1)

$m \backslash p$	1	2	3	4	5
1	1	2	3	4	5
2	2	3	6	7	8
3	3	6	8	9	11
4	4	7	9	10	17
5	5	8	11	17	19

LEMMA 3.4. $m + p - 1 \geq n$ is a sufficient condition for generic real pole assignment.

Proof: Consider the partition $\mu = (p^{m-1}, 1^p)$. As one immediately verifies, $K_\mu = 1$ corresponding to the only possibility:

$$\left(\begin{array}{ccccc} 1 & 1 & \dots & \dots & 1 \\ 2 & 2 & \dots & \dots & 2 \\ \vdots & & & & \vdots \\ m-1 & m-1 & \dots & \dots & m-1 \\ m & m+1 & \dots & \dots & m+p-1 \end{array} \right).$$

THEOREM 3.5. The following conditions imply generic real pole assignability. (By duality assume $m \leq p$) :

- (3.2)

$m = 2$ and $1.5p \geq n$
- (3.3)

$m = 3$ and $2p + 1 \geq n$
- (3.4)

$m \geq 4$ and $2.25p + m - 3 \geq n$.

The proof of this theorem and Theorem 2.3 is based on an interesting identification of the mod 2 cohomology ring of the real Grassmannian and the space of symmetric functions $\mathcal{Z}_2[x_1, \dots, x_p]^{\mathcal{S}_p}$. A good description of the topology of the real Grassmann manifold can be found for example in [14]. The important properties about the ring structure of $H^*(Grass(p, m+p), \mathcal{Z}_2)$ are given in the next section. Several properties about the ring of symmetric functions $\mathcal{Z}_2[x_1, \dots, x_p]^{\mathcal{S}_p}$ are summarized in an Appendix, where further references are given.

4. The cohomology ring of the real Grassmannian. The collection of m -planes in \mathbf{R}^{m+p} is called the Grassmann manifold and will be denoted by $Grass(p, m+p)$. The Grassmannian $Grass(p, m+p)$ is a smooth, compact manifold of dimension mp . Additively, the cohomology ring $H^*(Grass(p, m+p), \mathcal{Z}_2)$ can be described as a free \mathcal{Z}_2 -module over the set of Schubert cocycles $[a_1, \dots, a_p]$ where $m \geq a_1 \geq \dots \geq a_p \geq 0$. This notation coincides with the notation adopted in Griffith and Harris [6] and is “reverse” to the notation used by Hiller [7], [8].

Denote with $\xi_{p,m+p}$ the canonical p -bundle over $Grass(p, m+p)$. The total space of $\xi_{p,m+p}$ is defined by

$$(4.1) \quad E(\xi_{p,m+p}) = \{(V, x) \in Grass(p, m+p) \times \mathbf{R}^{m+p} \mid x \in V\}$$

and the corresponding bundle map is a projection on the first factor. The orthogonal bundle of $\xi_{p,m+p}$ is an m -plane bundle and will be denoted with $\bar{\xi}_{p,p+m}$. Finally denote with w_k the k th Stiefel Whitney class of $\xi_{p,p+m}$ and with σ_j the j th Stiefel Whitney class of $\bar{\xi}_{p,p+m}$. In terms of Schubert cocycles those Stiefel Whitney classes are described by

$$(4.2) \quad w_k = \underbrace{[1, 1, \dots, 1, 0, \dots, 0]}_k, \quad k = 1, \dots, p$$

$$(4.3) \quad \sigma_j = [j, 0, \dots, 0], \quad j = 1, \dots, m.$$

The multiplicative structure of $H^*(Grass(p, n), \mathcal{Z}_2)$ is described by the classical formulas of Pieri and Giambelli. Giambelli’s formula expresses a general Schubert cocycle as a polynomial in the special Schubert cocycle σ_j and Pieri’s formula explains how a Schubert cocycle is multiplied with a special Schubert cocycle.

Pieri’s formula:

$$(4.4) \quad [a_1, \dots, a_p] \cdot \sigma_j = \sum_{\substack{a_{i-1} \geq b_i \geq a_i \\ \sum_{i=1}^p b_i = (\sum_{i=1}^p a_i) + j}} [b_1, \dots, b_p]$$

Giambelli’s formula:

$$(4.5) \quad [a_1, \dots, a_p] = \det(\sigma_{a_i+j-i}) = \det \begin{pmatrix} \sigma_{a_1} & \sigma_{a_1+1} & \dots & \sigma_{a_1+p-1} \\ \sigma_{a_2-1} & \sigma_{a_2} & & \vdots \\ \vdots & & \ddots & \vdots \\ \sigma_{a_p-p+1} & & \dots & \sigma_{a_p} \end{pmatrix}.$$

From Giambelli’s formula it follows in particular that the Stiefel Whitney classes of the orthogonal bundle $\bar{\xi}_{p,p+m}$ generate $H^*(Grass(p, m+p), \mathcal{Z}_2)$. As shown by Hiller [8, p. 530] the same is true for the Stiefel Whitney classes of the canonical bundle $\xi_{p,p+m}$. In fact we will show that a general Schubert cocycle can be expressed in terms of the Stiefel-Whitney classes $\{w_1, \dots, w_p\}$ using a well-known classical formula.

In order to achieve our results the relation between the cohomology ring of the real Grassmannian and the space of symmetric functions will be studied. It will be shown that the cohomology ring $H^*(Grass(p, m+p), \mathcal{Z}_2)$ is isomorphic to a quotient of the space of symmetric functions $\mathcal{Z}_2[x_1, \dots, x_p]^{S_p}$. Using this identification it is possible to characterize the cup-length of $H^*(Grass(p, m+p), \mathcal{Z}_2)$ in a combinatorial way.

In the case of a complex Grassmannian a connection between the space of symmetric functions $\mathcal{Z}[x_1, \dots, x_p]^{S_p}$ and the cohomology ring $H^*(Grass(p, m+p), \mathcal{Z})$ is well known. According to Stanley [17], Lesieur [12] was the first who recognized a formal similarity between (4.5) and the classical identity of Jacobi and Trudi (see the Appendix). Horrocks [9] showed that this relationship is more than formal and can be explained geometrically.

In this section we work out a similar relationship for the real Grassmannian. From a geometric point of view, this relation can be understood in the following way.

Consider the space $Flag(\mathbf{R}^{m+p})$ of mutually orthogonal and ordered $(m+p)$ -tuples of lines (l_1, \dots, l_{m+p}) . Over $Flag(\mathbf{R}^{m+p})$ are line bundles ξ_i with total space $E(\xi_i)$, where

$$(4.6) \quad E(\xi_i) := \{((l_1, \dots, l_{m+p}); y) \in Flag(\mathbf{R}^{m+p}) \times \mathbf{R}^{m+p} \mid y \in l_i\}.$$

One has a projection

$$(4.7) \quad \begin{aligned} \pi : Flag(\mathbf{R}^{m+p}) &\longrightarrow Grass(p, m+p) \\ (l_1, \dots, l_{m+p}) &\longmapsto \text{span}(l_1, \dots, l_p) \end{aligned}$$

This projection induces an embedding (compare Hiller [8] or Stong [19])

$$(4.8) \quad \begin{aligned} \pi^* : H^*(Grass(p, m+p), \mathcal{Z}_2) &\longrightarrow H^*(Flag(\mathbf{R}^{m+p})) \\ &\cong \mathcal{Z}_2[x_1, \dots, x_{m+p}] / I_{mp}, \end{aligned}$$

where I_{mp} is the ideal generated by the relations

$$(4.9) \quad \prod_{i=1}^{m+p} (1 + x_i) = 1,$$

expressing the triviality of the bundle

$$\xi_1 \oplus \dots \oplus \xi_{m+p}.$$

The projection π can be covered by a bundle map. Indeed, consider the p -bundle $\xi_1 \times \dots \times \xi_p$ over $Flag(\mathbf{R}^{m+p})$. It is immediate that $\pi^*(w(\xi_{p, m+p})) = w(\xi_1 \times \dots \times \xi_p) = \prod_{i=1}^p (1 + x_i)$. Under π^* , the k th Stiefel Whitney class w_k of the canonical p -bundle $\xi_{p, p+m}$ of $Grass(p, p+m)$ is therefore mapped onto the k th elementary symmetric function $e_k = \sum x_{i_1} \dots x_{i_k}$ of $\mathcal{Z}_2[x_1, \dots, x_p]$.

Because the Schubert cocycles $\{w_1, \dots, w_p\}$ generate $H^*(Grass(p, m+p), \mathcal{Z}_2)$ as a ring, $H^*(Grass(p, m+p), \mathcal{Z}_2)$ can be embedded into $\mathcal{Z}_2[x_1, \dots, x_p]^{S_p} / \tilde{I}_{mp}$, where $\tilde{I}_{mp} = I_{mp} \cap \mathcal{Z}_2[x_1, \dots, x_p]^{S_p}$. Because the elementary symmetric functions generate $\mathcal{Z}_2[x_1, \dots, x_p]^{S_p}$, this last embedding is even an isomorphism.

In the following we will represent the ideal \tilde{I}_{mp} as the kernel of a ring homomorphism. For this denote with h_k the k th complete homogenous symmetric function in

p variables (see the Appendix for details). The set $B := \{h_1, \dots, h_p\}$ is algebraically independent and forms a multiplicative basis of $\mathcal{Z}_2[x_1, \dots, x_p]^{\mathcal{S}_p}$ (compare [13]). Any map defined on B extends therefore in a unique way to a ring homomorphism. Consider now the following ring homomorphism:

$$(4.10) \quad \begin{aligned} \psi : \mathcal{Z}_2[x_1, \dots, x_p]^{\mathcal{S}_p} &\longrightarrow H^*(Grass(p, m+p), \mathcal{Z}_2) \\ h_j &\longmapsto \sigma_j. \end{aligned}$$

Here we assume that the j th Stiefel Whitney class σ_j of the orthogonal bundle $\bar{\xi}_{p,p+m}$ is zero for $j > m$. Using the equivalence of (4.5) and the Jacobi–Trudi identity (6.10), it is immediate that a general Schur function s_λ is mapped onto the Schubert cocycle $[\lambda_1, \dots, \lambda_p]$. From Theorem 6.3 it follows that the k th elementary symmetric function e_k is equal to the Schur function $s_{(1^k, 0, \dots, 0)}$, and this element is mapped onto the Stiefel Whitney class w_k . Again from (4.5) it follows that the kernel of the map ψ has an additive basis of Schur functions s_λ with $\lambda_1 > m$. Finally the Nagelbasch–Kostka identity (6.11) gives a formula expressing a general Schubert cocycle as a polynomial in the Stiefel Whitney classes $\{w_1, \dots, w_p\}$.

5. Proof of the theorems. In the following denote by c the cup length of $H^*(Grass(p, m+p), \mathcal{Z}_2)$ and assume that $g = g_1 \cdots g_c$ is a maximal nonzero product. Our first goal will be to show that $g \in H^{mp}$, in other words, $g = [m^p]$. If not, expand g in terms of Schubert cocycles $g = \sum_{i \in I} [\lambda]^i$ and define $d := \max\{b \mid b = m - \lambda_p\}_{i \in I}$. From (4.4) it follows that $g \cdot \sigma_d \neq 0$ contradicting the maximality of the length of the product. It is therefore immediate that $d = 0$ and $g = [m^p]$.

Using the Nagelbasch–Kostka identity (6.11) one can express each factor g_i as a polynomial in the classes $\{w_1, \dots, w_p\}$. In this way, g becomes a polynomial $g = v(w_1, \dots, w_p)$. Because H^{mp} is one-dimensional, v is just a monom, in other words $g = w_\mu = w_{\mu_1} \cdots w_{\mu_k}$. During the substitution process, the number of factors can only increase, in other words $k \geq c$. On the other hand c is equal to the cup length, which shows $k = c$.

In $\mathcal{Z}_2[x_1, \dots, x_p]^{\mathcal{S}_p}$, this product corresponds to a product of elementary symmetric functions $e_\mu = e_{\mu_1} \cdots e_{\mu_c}$. To say w_μ is nonzero is therefore equivalent to the condition that $e_\mu \notin \ker(\psi)$. Using Theorem 6.3, one can expand e_μ in terms of Schur functions:

$$(5.1) \quad e_\mu = \sum_{|\lambda|=mp} K_{\bar{\lambda}\mu} s_\lambda.$$

Because $|\lambda| = |\mu| = mp$, there is exactly one Schur function s_λ not lying in the ideal $\tilde{I}_{mp} = \ker(\psi)$, namely, $s_\lambda = s_{(m^p)}$.

In summary, w_μ is nonzero if and only if the Kostka number $K_{(p^m)\mu}$ is odd. But this number is equal to the number K_μ introduced in §2. This proves Theorem 2.2 and therefore also Theorem 2.3. \square

In fact, one can show a little more. Using (4.5) and the same argument as above one finds a description of g in terms of the special Schubert cocycle σ_j , i.e., $g = \sigma_\nu = \sigma_{\nu_1} \cdots \sigma_{\nu_c}$. In $\mathcal{Z}_2[x_1, \dots, x_p]^{\mathcal{S}_p}$. This product can be written as:

$$(5.2) \quad h_\nu = h_{\nu_1} \cdots h_{\nu_c} = \sum_{|\lambda|=mp} K_{\lambda\nu} s_\lambda.$$

To say that σ_ν is nonzero is therefore equivalent to the condition that the Kostka number $K_{(m^p)\nu}$ is odd. In this way we have proved Lemma 5.1.

LEMMA 5.1. $c(m, p) = c(p, m)$.

The proof of Theorem 3.5 is partially based on results obtained by Stong [19]. In this paper Stong calculates explicitly maximal nonzero cup products $w_\mu = w_{\mu_1} \cdots w_{\mu_c}$. In this way he calculates the numbers $c(m, p)$ for $m = 2, 3, 4$.

Putting his results in a little more convenient form one obtains

$$(5.3) \quad c(2, p) = k_2(p) \cdot p \quad \text{where} \quad 1.5 \leq k_2(p) \leq 2,$$

$$(5.4) \quad c(3, p) = k_3(p) \cdot p + 1 \quad \text{where} \quad 2 \leq k_3(p) \leq 2.5,$$

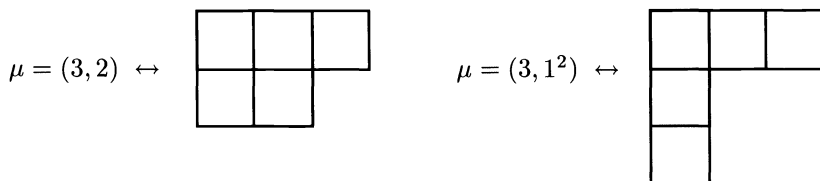
$$(5.5) \quad c(4, p) = k_4(p) \cdot p + 1 \quad \text{where} \quad 2.25 \leq k_4(p) \leq 3.$$

To get a lower bound for $c(m, p)$ in general ($m > 4$), Theorem 2.2 will be used. Consider a partition μ of the number $4p$ of length $c(4, p)$ in such a way that the Kostka number $K_{(p^4)(\mu)}$ is odd. But then it is immediate that the Kostka number $K_{(p^m)(p^{m-4}, \mu)}$ is odd as well. In this way one sees that $c(m, p) \geq c(4, p) + m - 4$, which completes the proof of Theorem 3.5. \square

6. Appendix: Symmetric functions. Let $\mu = (\mu_1, \dots, \mu_s)$ be a partition of n of length s . This means $\mu_1 \geq \mu_2 \geq \dots \geq \mu_s > 0$ and $\sum \mu_i = n$. If the integer μ_i is repeated r_i times in the partition μ , the abbreviated notation $\mu = (\mu_1^{r_1}, \dots, \mu_t^{r_t})$ will be used.

A partition μ defines a diagram D_μ , which can be considered as a left-justified array of boxes with μ_i boxes in the i th row.

Example 6.1. Two partitions with corresponding diagrams are illustrated:



The number $|\mu| = \sum \mu_i$ is sometimes called the *weight* of the partition μ and the numbers μ_i are called the *parts* of the partition. The dual partition $\bar{\mu} = (\bar{\mu}_1, \dots, \bar{\mu}_t)$ of a partition μ of n is obtained by taking the “transpose” of D_μ . In other words $\bar{\mu}_i$ is defined as the number of boxes in the column i of D_μ . Assume in addition that there is given a set $S \subseteq \mathcal{N}$.

DEFINITION 6.2. A standard Young tableau of shape μ is a diagram D_μ , where each box in D_μ contains a number from S under the constraint that the rows are increasing and the columns are strictly increasing.

Consider now $R = \mathcal{Z}_2[x_1, \dots, x_p]^{S_p}$, the ring of symmetric functions in p variables. R is in a natural way a graded ring:

$$(6.1) \quad R = A_0 + A_1 + \dots + A_n + \dots, \quad A_i A_j \subseteq A_{i+j}.$$

The homogenous component A_n can be described by different classical bases, where each basis is usually parametrized by the set of all partitions μ with weight n . In particular the dimension of A_n is equal to $p(n)$, the partition number of n .

Products of elementary symmetric functions:

$$(6.2) \quad e_\mu := e_{\mu_1} \cdots e_{\mu_s},$$

where $e_k = \sum x_{i_1} \cdots x_{i_k}$ is the k th elementary symmetric function.

Monomial symmetric functions:

$$(6.3) \quad m_\mu := \sum x_1^{\mu_1} \cdots x_p^{\mu_p},$$

where the summation has to be taken over all distinct monomials with exponents μ_1, \dots, μ_p .

Complete homogenous symmetric functions:

$$(6.4) \quad h_\mu := h_{\mu_1} \cdots h_{\mu_s},$$

where $h_k = \sum_{|\lambda|=k} m_\lambda$ is the k th complete symmetric function.

Schur functions: Classically, Schur functions were introduced by Jacobi (~1835) as the quotient of two alternating functions giving a symmetric function:

$$(6.5) \quad s_\mu = \frac{\det[x_i^{\mu_j + p - j}]}{\det[x_i^{p-j}]}, \quad i, j = 1, \dots, p.$$

The denominator of this expression is nothing else than the Vandermonde determinant and the numerator is a generalization of this type of determinant. The importance of those functions became apparent when Schur, a student of Frobenius, developed the character theory of the symmetric group (~1900).

The change of basis between the different bases of A_n is described by a linear transformation. In 1907, Kostka [11] published matrices describing the change of basis and showed that the different transformations are closely related. The following theorem, which is due to Kostka, is proven in [18].

THEOREM 6.3.

$$(6.6) \quad h_\mu = \sum_{|\lambda|=n} K_{\lambda\mu} s_\lambda,$$

$$(6.7) \quad s_\mu = \sum_{|\lambda|=n} K_{\mu\lambda} m_\lambda,$$

$$(6.8) \quad e_\mu = \sum_{|\lambda|=n} K_{\bar{\lambda}\mu} s_\lambda.$$

The coefficients $\{K_{\lambda\mu}\}$ are called the *Kostka coefficients*. The number $K_{\lambda\mu}$ can be described in a combinatorial way as the number of standard Young tableaux with shape λ and content μ . This means the number of ways of filling in μ_i integers i into the diagram D_λ under the condition that the rows are increasing and the columns are strictly increasing (see [18]).

Only in very special cases are formulas for the numbers $K_{\lambda\mu}$ known. For example if $\mu = (1, 1, \dots, 1)$, the number $K_{\lambda\mu}$ can be described by the famous hook formula of Frame, Robinson, and Thrall [5]. If in addition $\lambda = (p^m)$, the number $K_{\lambda\mu}$ is described by the following formula:

$$(6.9) \quad K_{(p^m)(1^mp)} = \frac{1! \cdots (p-1)! \cdot (mp)!}{m! \cdots (m+p-1)!}.$$

Finally, the following two classical formulas give a polynomial expression of a Schur function in terms of complete symmetric respective elementary symmetric functions.

Jacobi–Trudi identity:

$$(6.10) \quad s_\lambda = \det(h_{\lambda_i+j-i}), \quad i, j = 1, \dots, p.$$

Nagelbasch–Kostka identity:

$$(6.11) \quad s_\lambda = \det(e_{\bar{\lambda}_i+j-i}), \quad i, j = 1, \dots, \lambda_1.$$

More details about these identities are given in [13, p. 25] and in [18].

Acknowledgments. I would like to thank my thesis adviser and teacher Chris Byrnes for introducing me to the pole placement problem and for his constant encouragement during my graduate studies. For helpful comments I thank Stephan Stolz and the anonymous referees.

REFERENCES

- [1] R. W. BROCKETT AND C. I. BYRNES, *Multivariable Nyquist criteria, root loci and pole placement: A geometric viewpoint*, IEEE Trans. Automat. Control, AC-26, (1981), pp. 271–284.
- [2] C. I. BYRNES, *On the stability of multivariable systems and the Ljusternik–Snirelmann category*, Systems Control Lett. 3, (1983), pp. 255–262.
- [3] ———, *Pole assignment by output feedback*, Lecture Notes in Control and Inform. Sciences, No. 135, Springer–Verlag, Berlin, Heidelberg, New York, 1989, pp. 31–78.
- [4] S. EILENBERG, *Sur un théorème topologique de M.L.Snirelmann*, Mat. Sb., 1 (1936), pp. 557–559.
- [5] J. S. FRAME, G. B. ROBINSON AND R. M. THRALL, *The hook length of S_n* , Canad. J. Math., 6 (1954), pp. 316–325.
- [6] P. GRIFFITH AND J. HARRIS, *Principles of Algebraic Geometry*, John Wiley and Sons, New York, 1978.
- [7] H. HILLER, *Geometry of Coxeter Groups*, Res. Notes Math., 54, Pitman, 1982.
- [8] ———, *On the cohomology of real grassmannian*, Trans. Amer. Math. Soc., 79 (1980), pp. 521–533.
- [9] G. HORROCKS, *On the relations of S -functions to Schubert varieties*, Proc. London Math. Soc., (3), 7 (1957), pp. 265–280.
- [10] H. KIMURA, *Pole assignment by gain output feedback*, IEEE Trans. Automat. Control, 20 (1975), pp. 509–516.
- [11] C. KOSTKA, *Tafeln und Formeln für symmetrische Funktionen*, Jahresbr. Deutsch. Math. Verein., 16 (1907), pp. 429–450.
- [12] L. LESIEUR, *Les problèmes d’intersection sur une variété de Grassmann*, C.R. Acad. Sc. Paris, 225 (1947), pp. 916–917.
- [13] I. G. MACDONALD, *Symmetric Functions and Hall Polynomials*, Oxford University Press, Oxford, 1979.
- [14] J. W. MILNOR AND J. D. STASHEFF, *Characteristic Classes*, Ann. Math. Stud., No. 76, Princeton, NJ, 1974.
- [15] J. ROSENTHAL, *Geometric Methods for Feedback Stabilization of Multivariable Linear Systems*, Ph.D. thesis, Arizona State University, 1990.
- [16] ———, *Tuning natural frequencies by output feedback*, in Computation and Control, K. Bowers and J. Lund, eds., Birkhäuser, Boston, 1989, pp. 276–282.
- [17] R. STANLEY, *Some combinatorial aspects of the Schubert calculus*, Lecture Notes in Math., No. 579, Springer–Verlag, Berlin, New York, 1977, pp. 217–251.
- [18] ———, *Theory and applications of plane partitions I, II*, Stud. Appl. Math., 50 (1971), pp. 167–188, 259–279.
- [19] R. E. STONG, *Cup products in Grassmannians*, Topology Appl. 13 (1982), pp. 103–113.
- [20] X. WANG, *On output feedback via Grassmannians*, SIAM J. Control Optim., 29 (1991), pp. 926–935.
- [21] J. C. WILLEMS AND W. H. HESSELINK, *Generic properties of the pole placement problem*, Proc. IFAC, Helsinki, Finland, 1978.

INFORMATION AND STRATEGIES IN DYNAMIC GAMES*

P. BERNHARD†

Abstract. This paper extends to the setting of stochastic dynamic games with incomplete information a theorem of Kuhn and uses it to prove the existence of a saddle point in a suitable class of strategies. It then particularizes this result to the situation where one of the players has full information to show existence of a saddle point in another class of strategies exhibiting a constant dimension sufficient statistic. A dynamic programming-like algorithm is naturally associated with this class of strategies and was proposed in a previous paper in a sufficient condition setting. For this same class of games, an example of another use of the main theorem is given, leading to a different dynamic programming-like algorithm.

Key words. games, saddle point, mixed strategies, dynamic programming

AMS(MOS) subject classifications. 90D20

1. Introduction. In an historical paper in 1953 [5], Kuhn introduced the modern concept of game in extensive form, extending and simplifying the concept introduced by von Neuman and Morgenstern [11]. In the same paper, he proved that all such games *with complete memory* have a saddle point in behavioral strategies.

We shall consider the equivalent property in the setting of dynamic, and more specifically, multistage games. When all variables range over finite sets, the latter are a special case of games in extensive form. However, our approach allows us to deal with the case where the decision variables range over infinite sets (we shall restrict them to compact sets for technical reasons) and with noisy information. Notice also that the property of *perfect memory*, which was somewhat technical in Kuhn's setup, becomes extremely natural and simple in the setup of dynamical systems.

For the sake of simplicity, we restrict our attention to two-player zero-sum games. It is clear that our form of Kuhn's theorem, as well as the original one, extends to many-person games. (The simplest way to see that is, following Aumann [1], to lump into "player two" the actions of all the other players.) Again for simplicity, we shall first derive it for deterministic games and extend it to stochastic games after. Of course, as in refs [8], [9] discussed below, we also could assume compacity of the control space of one player only and use the nonsymmetric version of Sion's theorem.

In his paper [1], Aumann proposed an extension of Kuhn's theorem to infinite games. He states (p. 628), "A mixed strategy can be thought of as a probability distribution, i.e., a measure, on the set of all pure strategies." This is exactly what we do here. He prefers not to place a measurable structure on the set of pure strategies. (We use Radon measures, with various topologies.) As a result, he has to use a rather restrictive class of mixed strategies, which are not really the direct generalization of Kuhn's mixed strategies but some superset of our, and his, behavioral strategies. But the difference between mixed and behavioral then appears somewhat artificial, since his definition of behavioral is that, in our notations, for $i > j$, the random variables u_i and u_j should be independent. However, the probability law of u_i is

*Received by the editors August 14, 1989; accepted for publication (in revised form) January 25, 1991.

†Institut National de Recherche en Informatique et en Automatique, Sophia Antipolis, Valbonne, France.

$p_i = \varphi_i(u^{t-1}, y^t)$, explicitly depending on the realization u_j . (In the notations of [1], our u_i is $y_i = b_i(x_i, \omega)$, where x_i plays the role of our r_i , and encodes, through the function u_i^j , the information on $y_j = u_i^j(x_i)$.)

In the past few years, several papers have dealt with this type of game, see, e.g., [8] and [9]. These two references and the related literature use a similar set up to ours, and their strategies are our behavioral strategies. The link with what we call mixed strategies is not made there. They do not consider either deterministic games (which are topologically more difficult to handle), nor, more importantly, partial or noisy state information. They deal with infinite time games. The main obstacle to doing so here is that the second part of the paper would not carry over in a simple way.

Starting with §7, we examine in more detail the case where one of the players has full (causal) information, where it is known that the second guessing problem simplifies. See [2] and the bibliography therein. A dynamic programming approach lets us achieve two things. On the one hand, it allows us to show that there exists a saddle point in the class of strategies used in [2], using finite dimensional sufficient statistics, which turns the sufficient condition of that paper into a necessary and sufficient condition. On the other hand, it may be more effective for short duration games to stay with the space of behavioral strategies, and we shall derive a different dynamic programming-like algorithm for the particular case of the rabbit and hunter game.

NOTATION. We shall study only *discrete time* games, so that we shall have to deal with finite sequences of objects. We shall adopt the following conventions. Let $a = \{a_1, a_2, \dots, a_T\}$ be a finite sequence, where $a_t \in A_t$. A subscript will refer to a particular element of the sequence, while a superscript will refer to the restriction of the sequence to its first elements: $a^t = \{a_1, a_2, \dots, a_t\} \in A_1 \times A_2 \times \dots \times A_t = A^t$. The notation A^t will therefore mean the cartesian power of A *only if* $A_1 = A_2 = \dots = A_t = A$. Likewise, if α is a function ranging over A^T , α_t will be its component in A_t , and α^t its first t components. Finally, if $a \in A^t$ and $b \in A_{t+1}$, then $a \cdot b$ stands for the element of A^{t+1} obtained by concatenating a and b .

Let us also agree that for a topological space A , we shall call $\pi(A)$ the set of all (Radon) probability measures over A .

2. Multistage game. A deterministic two-player zero-sum multistage game is given by

- An integer T called the *horizon* of the game. Let $\mathbf{T} = \{1, 2, \dots, T\}$ and $t \in \mathbf{T}$ is called the time.
- A sequence of *state spaces* X_t . We shall use $x_t \in X_t$, the *state at time* t .
- An *initial state* $x_1 \in X_1$, which is assumed to be part of the common knowledge of both players.
- Two sequences of *output spaces* Y_t and Z_t . $y_t \in Y_t$ and $z_t \in Z_t$ are the *measurements* of player 1 and 2 at time t .
- Two *control sets* \mathcal{U} and \mathcal{V} and two point to set maps *admissible controls* $y_t \mapsto U(y_t) \subset \mathcal{U}$ and $z_t \mapsto V(z_t) \subset \mathcal{V}$. We shall oftentimes write U_t and V_t instead of $U(x_t)$ and $V(x_t)$ when what is meant is clear. $u_t \in U_t$ and $v_t \in V_t$ are the *controls* of player 1 and 2 respectively at time t .
- A sequence of functions *dynamics* $f_t : X_t \times U_t \times V_t \rightarrow X_{t+1}$.

$$(1) \quad x_{t+1} = f_t(x_t, u_t, v_t).$$

- Two sequences of *output functions* $h_t : X_t \rightarrow Y_t$ and $k_t : X_t \rightarrow Z_t$, defining

$$(2a) \quad y_t = h_t(x_t),$$

$$(2b) \quad z_t = k_t(x_t),$$

(Actually, h_t and k_t need only be defined for $t \geq 2$)

- A *Capture set* $C \subset X^T \times \mathbf{T}$ defining the *final time* t_1 through

$$(3) \quad t_1 = \begin{cases} \min\{t \mid (x_t, t) \in C\} & \text{if capture happens,} \\ T & \text{otherwise.} \end{cases}$$

- A *criterion* (or *cost function*) G that player **1**, choosing the control, $u_t \in U_t$ strives to minimize, and **2**, choosing the controls $v_t \in V_t$ to maximize. G is defined via two sequences of functions: $L_t : X_t \times \mathcal{U} \times \mathcal{V} \rightarrow \mathbf{R}$ and $K_t : X_t \rightarrow \mathbf{R}$ by

$$(4) \quad G = \sum_{t=1}^{t_1-1} L_t(x_t, u_t, v_t) + K_{t_1}(x_{t_1}).$$

Remark. The above setup is that of dynamical systems and is by now classical. The important point for us is that it defines functions

$$(5a) \quad y_t = \tilde{h}_t(u^{t-1}, v^{t-1}),$$

$$(5b) \quad z_t = \tilde{k}_t(u^{t-1}, v^{t-1}).$$

We shall also make use of

$$(6a) \quad y^t = h^t(x^t) = \tilde{h}^t(u^{t-1}, v^{t-1}),$$

$$(6b) \quad z^t = k^t(x^t) = \tilde{k}^t(u^{t-1}, v^{t-1}),$$

and also

$$(7) \quad G = \tilde{G}_{t_1}(u^{t_1-1}, v^{t_1-1}).$$

Let us introduce a last notation. We shall be interested in games with complete memory. By this we mean that the *information* available to the players to make up their choice of control values at each instant of time is the whole sequence of their own past controls and their past and present measurements. We shall therefore write:

$$(8a) \quad r^t = (u^{t-1}, y^t) \quad \text{player 1's information,}$$

and

$$(8b) \quad s^t = (v^{t-1}, z^t) \quad \text{player 2's information.}$$

We shall also use the following definition

DEFINITION. If the sets \mathcal{U} and \mathcal{V} , (and for stochastic games, \mathcal{W}), are all finite, the game shall be called *finite*.

Note that for a finite game, the sets X_t , Y_t , and Z_t may, with no loss of generality be restricted to finite sets.

For infinite games, we shall use the following *topological hypothesis*.

HYPOTHESIS. The sets U_t , V_t , X_t , Y_t , and Z_t are all topological spaces, the first two *compact*. The functions f_t , h_t , k_t , L_t , and K_t are all continuous, (making G in (4) a continuous function of (x^T, u^{T-1}, v^{T-1}) , or \tilde{G} in (7) a continuous function of (u^{T-1}, v^{T-1})).

3. Strategies. To make precise the definition of the game, we must now specify the strategy sets and what are the quantities to be minimized or maximized.

DEFINITION. We call *pure strategy* of player **1** a *non anticipatory* measurable map

$$(9a) \quad \alpha : Y^{T-1} \rightarrow U^{T-1} : (y_1, \dots, y_{T-1}) \mapsto (u_1, \dots, u_{T-1})$$

and we call A the set of all such nonanticipatory maps, i.e., such that, for a and b in Y^{T-1}

$$a^t = b^t \implies \alpha_t(a) = \alpha_t(b).$$

Likewise, pure strategies of the second player are nonanticipatory measurable maps

$$(9b) \quad \beta \in B, \quad \beta : Z^{T-1} \rightarrow V^{T-1} : (z_1, \dots, z_{T-1}) \mapsto (v_1, \dots, v_{T-1}).$$

Any pair of pure strategies $(\alpha, \beta) \in A \times B$ generates a well defined game history by (1), (2), and

$$(10a) \quad u_t = \alpha_t(y_1, \dots, y_t, \eta),$$

$$(10b) \quad v_t = \beta_t(z_1, \dots, z_t, \zeta),$$

where η and ζ are arbitrary sequences that do not affect the resulting values of u_t and v_t , by the hypothesis of nonanticipativity. As a consequence, we shall omit them in the future. There corresponds to it a well-defined cost

$$G = \hat{G}(\alpha, \beta).$$

We can state the following fact.

LEMMA. *Under the topological hypothesis, the sets A and B of pure strategies are compact in the topology of pointwise convergence.*

Proof. The set of all functions from Y^{T-1} into U^{T-1} is isomorphic to the power set

$$(U^{T-1})^{Y^{T-1}}.$$

Since the U_t are all assumed compact, by Tychonov's theorem so is U^{T-1} , and therefore also the above power set, in the product topology, which coincides with the topology of pointwise convergence. Now, the property of nonanticipativity is clearly preserved in the pointwise limit, so that the sets of pure strategies are closed subsets of compact sets in that topology. \square

In the case where \hat{G} has no saddle point over $A \times B$, it is natural to introduce mixed strategies. Assuming we have chosen a topology on A and B , we may give the following definition.

DEFINITION. We call *mixed strategies* probability measures λ and μ over A and B respectively:

$$(11a) \quad \lambda \in \pi(A),$$

$$(11b) \quad \mu \in \pi(B).$$

We then wish to take as a new criterion

$$(12) \quad J(\lambda, \mu) = E(G) = \int_{A \times B} \hat{G}(\alpha, \beta) d\lambda(\alpha) d\mu(\beta).$$

The existence of the above integral is not guaranteed a priori, since \hat{G} is in general not continuous in α and β , and there is no guarantee that it be measurable. Three alternate sets of hypotheses are provided here as examples of setups where this integral exists and that preserve the compactness of the set of allowed pure strategies. The first set is rather trivial.

HYPOTHESIS 1. The game is finite.

Then the integral in (12) is merely a finite sum. The game is just a matrix game, (possibly with a very large matrix!), and we are in the classical setting of von Neuman and Morgenstern.

The second set allows us to avoid any finiteness hypothesis, but imposes rather strong restrictions on the allowed pure strategies.

HYPOTHESIS 2. The sets U_t , V_t , Y_t , and Z_t , are metric compact spaces, the pure strategies are restricted to be Lipschitz continuous with a prescribed Lipschitz modulus.

We then have the following fact.

LEMMA. *Under Hypothesis 2, the sets A and B are compact in the topology of uniform convergence, and the sequences u^{T-1} , v^{T-1} , and x^T depend continuously on (α, β) .*

Proof. The first claim is a direct consequence of the Arzela Ascoli theorem. The second claim derives from the fact that the x_t 's are then continuous functions of the strategies, as is easily seen by induction on t . \square

Finally, it is also possible to avoid regularity assumptions on the pure strategies, still keeping infinite control sets, by assuming finite observation sets. Notice that it follows from the standing topological hypothesis that the reachable game space is bounded, so that any quantization of the measurement with a finite mesh will produce a finite measurement set. This type of hypothesis was first proposed by Levine [6].

HYPOTHESIS 3. The sets Y_t and Z_t are finite. The functions h_t and k_t can therefore not be continuous. Assume that the sets $h^{-1}(y_k) \cap k^{-1}(z_l)$ have nonvoid interiors, the union of their boundaries is made of the finite union of sets on which h and k are constant, that satisfy the same hypothesis in the relative topology of this boundary, and so on recursively, the whole construction defining a finite partition of X^T .

LEMMA. *Under Hypothesis 3, the set $A \times B$ can be partitioned in a finite union of Borel sets (in the topology of pointwise convergence), the state trajectory x^T depending continuously on (α, β) over each of them, and thus also the control sequences u^{T-1} , v^{T-1} .*

Proof. Let, for simplicity, $A \times B = C$, $(\alpha, \beta) = \gamma$, and, for this proof, y stand for (y, z) . Let also $P_k, k = 1, \dots, K$ be the interior of the sets in X such that $y = \text{const}$, $P = \cup P_k$, and

$$C_t = \{\gamma \mid x_s \in P, \forall s \leq t\}.$$

It is easy to see by induction on t that C_t is open in C . As a matter of fact, let $\gamma^{(n)}$ be a sequence of strategies converging to $\gamma \in C_t$. Since y_1 is fixed, $u_1^{(n)} = \gamma_1^{(n)}(y_1) \rightarrow u_1$. By continuity of f_1 , $x_2^{(n)} \rightarrow x_2$. Since x_2 is interior to $h_2^{-1}(y_2)$, for n large enough, $y_2^{(n)} = y_2$. Therefore, we can use the convergence of $\gamma_2^{(n)}$ and the continuity of f_2 to

conclude that $x_3^{(n)} \rightarrow x_3$, and so on. Therefore $x_t^{(n)} \rightarrow x_t$, and for n large enough, $x_t^{(n)}$ is also in the interior of its set P_k , thus $\gamma^{(n)}$ is in C_t .

Therefore, the complement D_t of C_t in C_{t-1} is also a Borel set. This is the set of strategies γ such that $x_s \in P$, for $s = 1, \dots, t-1$, $x_t \in \partial P$. Now, assume again that $x_t \in Q_i$, where Q_i is the relative interior of one of the subsets of constant y in ∂P . The same type of argument will show that the subset E_l of D_t such that the next x_s , up to $s = l$, are in P is again open in D_t .

We finally have a finite number of subsets of C , depending on the subsets of the partition of X in which each of the x_t lie. All are Borel sets. And since we have shown the convergence of the $x_t^{(n)}$ in each case, \hat{G} is continuous in the relative interior of all of them. This proves the lemma. \square

As a consequence, \hat{G} is measurable, and J in (12) is well defined.

So we now have at least three cases where the following *existence hypothesis* is satisfied.

HYPOTHESIS. For a suitable topology on A and B , these sets are compact, and the state trajectory x^T , as well as the control histories u^{T-1} , v^{T-1} , are measurable functions of the pure strategies α and β .

As a consequence of this hypothesis, we have the following two facts.

PROPOSITION 0. *J in (12) is well defined, since G as defined in (4) is a continuous function of (x^T, u^{T-1}, v^{T-1}) .*

THEOREM 0. *Under the existence hypothesis, and if furthermore, \hat{G} is continuous, the game has a saddle point in mixed strategies.*

Proof. See Ekeland [4, p. 25] The proof makes use of Sion's theorem (see [10]) together with the vague topology on the set of measures.

The idea behind mixed strategies is that the players choose a pure strategy at random, according to the probability laws λ and μ respectively, once for all at the beginning of the game and for its whole duration. The spaces of mixed strategies are very large and complicated sets, and this type of behavior may not appear very natural. We shall therefore introduce another concept of strategies. But we first need a preliminary definition.

DEFINITION. We call *mixed control* of the first player at time t a probability law p_t over U_t , and likewise for the second player.

Let thus

$$(13a) \quad p_t \in P_t = \pi(U_t),$$

$$(13b) \quad q_t \in Q_t = \pi(V_t).$$

We now define the new class of strategies.

DEFINITION. We call *behavioral strategy* of the first player a sequence of measurable maps

$$(14a) \quad \varphi_t \in \Phi_t, \quad \varphi_t : U^{t-1} \times Y^t \rightarrow P_t : (u^{t-1}, y^t) \mapsto p_t = \varphi_t(u^{t-1}, y^t) = \varphi_t(r^t),$$

and similarly for the second player:

$$(14b) \quad \psi_t \in \Psi_t, \quad \psi_t : V^{t-1} \times Z^t \rightarrow Q_t : (v^{t-1}, z^t) \mapsto q_t = \psi_t(v^{t-1}, z^t) = \psi_t(s^t).$$

The game is then extended by considering $\{x_t\}$, $\{u_t\}$, $\{v_t\}$, $\{y_t\}$, and $\{z_t\}$ as stochastic processes, generated by (1) and (2), u_t and v_t being stochastic variables

with probability distributions p_t and q_t respectively, given by (14). Then the sequences u^{T-1} and v^{T-1} are stochastic variables, and we define the payoff of the game as

$$(15) \quad J = E(G) = E\left(\tilde{G}(u^{T-1}, v^{T-1})\right).$$

This is well defined, since \tilde{G} is continuous, and (14) clearly defines a Radon probability over $U^{T-1} \times V^{T-1}$. (By a finite, elementary, version of the Ionescu Tulcea theorem.)

The idea behind behavioral strategies is that at each instant of time, the players choose their controls at random, according to a probability distribution function of their information.

One might believe that this new set of strategies is richer than the previous mixed strategies, since it involves many random choices instead of a single one. A very simple example will teach us that this is not so.

4. Example. The following example is the smallest possible version of the Hunter and Rabbit game. Let $T = 4$ (i.e., the game has three time steps.) The state is $x = (y, w, z) \in \{1, 2\} \times \{0, 1, 2\}^2$ and $U = V = \{1, 2\}$. The dynamics are

$$\begin{aligned} y_{t+1} &= u_t, & y_1 &= 1, \\ w_{t+1} &= v_t, & w_1 &= 0, \\ z_{t+1} &= w_t, & z_1 &= 0. \end{aligned}$$

Player 1 has no other information than the sequence u^{t-1} at time t . Player 2 knows the whole state in addition to the past controls. The capture set is $y_t - z_t = 0$. The payoff to player 2 is 1 if $y_{t_1} = z_{t_1}$, i.e., if *capture* has occurred, and 0 otherwise. Hence, $J = E(G)$ is the capture probability. Recall that we always assume that the initial state is part of the rule of the game, and is therefore common knowledge.

Player 1 actually plays open loop, and therefore chooses among $2^3 = 8$ pure strategies.

Let us look at the possible strategies of player 2. The pure strategies are specified by the decision rules at time 1 and 2, since actions taken at time 3 have no effect on the outcome of the game. At time 1, no information is available beyond the rules of the game. The only two possibilities are either $v_1 = 1$ or $v_1 = 2$. At time 2, a measurement $y_2 = u_1$ is available, which can be equal to 1 or 2. Since v_2 can also be chosen as either 1 or 2, there are 4 possible decision rules for $\beta_2(y_2)$, i.e., $v_2 = 1 \forall y$, that we denote by **1**, or, with the same convention, **2**, or $v_2 = y_2$, or finally $v_2 = 3 - y_2$. We thus have the following list of 8 possible pure strategies:

strategy	1	2	3	4	5	6	7	8
time								
$t = 1$	1	1	1	1	2	2	2	2
$t = 2$	1	2	y_2	$3 - y_2$	1	2	y_2	$3 - y_2$

The mixed strategies are therefore given by 8 probabilities (μ_1, \dots, μ_8) whose sum is one, thus seven degrees of freedom.

Let us now look at the behavioral strategies, still for player 2. They are defined by the probability ψ_1 of playing $v_1 = 1$, (and $1 - \psi_1$ of playing $v_1 = 2$), and for time $t = 2$ the four probabilities $\psi_2(s^2)$ of playing $v_2 = 1$ according to the four possible values of $s^2 = (v_1, y_2)$. Thus, these strategies are defined by five independent probabilities.

We also see on this example that one can make a mixed strategy correspond to a unique behavioral strategy, its *behavior*, by seeing the latter as a conditional marginal probability. Here, for instance, assuming that the pure strategies are numbered according to the above table, we have

$$\begin{aligned}\psi_1 &= \Pr(v_1 = 1) = \mu_1 + \mu_2 + \mu_3 + \mu_4, \\ \psi_2(1, 1) &= \Pr(v_2 = 1 | v_1 = 1, y_2 = 1) = \frac{\mu_1 + \mu_3}{\mu_1 + \mu_2 + \mu_3 + \mu_4} \\ \psi_2(1, 2) &= \Pr(v_2 = 1 | v_1 = 1, y_2 = 2) = \frac{\mu_1 + \mu_4}{\mu_1 + \mu_2 + \mu_3 + \mu_4}\end{aligned}$$

Likewise, the conditioning on $v_1 = 2$ would lead to a denominator $\mu_5 + \mu_6 + \mu_7 + \mu_8$, and so on. It is easy to verify that this map is onto, but cannot be one-to-one. An infinity of different mixed strategies lead to the same behavioral strategy.

The aim of the next section of this paper is to show that it is enough to consider behavioral strategies, for the outcome of the game only depends on the behavior of the strategies used, as rigorously defined hereafter.

Let us remark before we close this example that the solution of this game is almost obvious: each player should play either 1 or 2 with probability $\frac{1}{2}$.

5. Kuhn's theorem. We must first make precise the relationship between a mixed strategy and its associated behavioral strategy.

DEFINITION. We define the map *behavior* γ from the set $\pi(A)$ of mixed strategies to the set Φ of behavioral strategies in the following way:

$\varphi_t(u^{t-1}, y^t)$ is the marginal law on $\alpha_t(y^t)$ knowing $\alpha^{t-1}(y^{t-1}) = u^{t-1}$.

In the finite case, for instance, this can be made explicit in the following way. Let y^t and u^{t-1} be fixed. Let

$$\begin{aligned}A_1 &= \{\alpha \in A | \alpha^{t-1}(y^{t-1}) = u^{t-1}\}, \\ A_2 &= \{\alpha \in A | \alpha^t(y^t) = u^{t-1} \cdot \bar{u}\} \subset A_1.\end{aligned}$$

Then, $\gamma(\lambda) = \varphi$ with, if $A_1 \neq \emptyset$

$$\varphi_t[u^{t-1}, y^t](\bar{u}) = \left(\sum_{\alpha \in A_2} \lambda(\alpha) \right) \left(\sum_{\alpha \in A_1} \lambda(\alpha) \right)^{-1}.$$

If $A_1 = \emptyset$, $\varphi_t[u^{t-1}, y^t]$ may be arbitrarily specified, and we shall always assume that $\varphi = \gamma(\alpha)$ has been so extended to all values of its arguments.

We shall also call γ the behavior map of the second player, from $\pi(B)$ into Ψ .

We now state the main theorem, which is an extension of Kuhn [5]:

THEOREM 1. *There exists a function \bar{J} from $\Phi \times \Psi$ into \mathbf{R} such that,*

$$\forall(\lambda, \mu) \in \pi(A) \times \pi(B), \quad J(\lambda, \mu) = \bar{J}(\gamma(\lambda), \gamma(\mu))$$

i.e., the criterion $J(\lambda, \mu)$ of the game only depends on the behaviors $\gamma(\lambda)$ and $\gamma(\mu)$.

Proof. Under the existence hypothesis, (u^{T-1}, v^{T-1}) is a measurable function of (α, β) . Thus a pair of mixed strategies (λ, μ) generates a probability distribution Π^{T-1} over the set $U^{T-1} \times V^{T-1}$, and one has

$$J(\lambda, \mu) = \int_{U^{T-1} \times V^{T-1}} \tilde{G}(u^{T-1}, v^{T-1}) d\Pi^{T-1}(u^{T-1}, v^{T-1}),$$

or, in the discrete case

$$J(\lambda, \mu) = \sum_{u^{T-1}, v^{T-1}} \tilde{G}(u^{T-1}, v^{T-1}) \Pi^{T-1}(u^{T-1}, v^{T-1}).$$

Therefore, J only depends on the probability law Π^{T-1} . Notice that, for fixed $\bar{u}^{T-1}, \bar{v}^{T-1}$, the event $(u^{T-1}, v^{T-1}) = (\bar{u}^{T-1}, \bar{v}^{T-1})$ can be written

$$(u^{T-2}, v^{T-2}) = (\bar{u}^{T-2}, \bar{v}^{T-2}) \quad \text{and} \quad (u_{T-1}, v_{T-1}) = (\bar{u}_{T-1}, \bar{v}_{T-1}).$$

Furthermore, it follows from the definition (12) that the strategies λ and μ are chosen independently by the two players, so that, for $u^{T-2} = \bar{u}^{T-2}$ and $v^{T-2} = \bar{v}^{T-2}$ fixed, the random variables u_{T-1} and v_{T-1} are independent.

Let therefore λ and μ be fixed, $\varphi = \gamma(\lambda)$ and $\psi = \gamma(\mu)$ their behaviors,

$$\begin{aligned} \tilde{h}^{T-1}(\bar{u}^{T-2}, \bar{v}^{T-2}) &= \bar{y}^{T-1}, & \tilde{k}^{T-1}(\bar{u}^{T-2}, \bar{v}^{T-2}) &= \bar{z}^{T-1}, \\ \varphi_{T-1}(\bar{u}^{T-2}, \bar{y}^{T-1}) &= \bar{p}_{T-1}, & \psi_{T-1}(\bar{v}^{T-2}, \bar{z}^{T-1}) &= \bar{q}_{T-1}. \end{aligned}$$

One has the equality

$$d\Pi^{T-1}(\bar{u}^{T-1}, \bar{v}^{T-1}) = d\Pi^{T-2}(\bar{u}^{T-2}, \bar{v}^{T-2}) d\bar{p}_{T-1}(\bar{u}^{T-1}) d\bar{q}_{T-1}(\bar{v}^{T-1}).$$

(See, for instance, the second part of proposition V.1.1 in [7].)

Or in the discrete case,

$$\Pi^{T-1}(\bar{u}^{T-1}, \bar{v}^{T-1}) = \Pi^{T-2}(\bar{u}^{T-2}, \bar{v}^{T-2}) \bar{p}_{T-1}(\bar{u}_{T-1}) \bar{q}_{T-1}(\bar{v}_{T-1}),$$

by Bayes rule.

One then iterates this process to write Π^{T-2} in terms of Π^{T-3} and of φ_{T-2} and ψ_{T-2} , and so on. This proves the theorem. \square

This allows us to carry over results obtained by topological means on games in normal form, such as Theorem 0 above, to dynamic games in behavioral strategies. We thus have, for instance, the following obvious fact.

COROLLARY 1. *Under the existence hypothesis, and if furthermore \hat{G} is continuous, a dynamic game with perfect memory admits a saddle point in behavioral strategies over the sets $\Phi = \gamma(A)$, $\Psi = \gamma(B)$.*

The difficult task, however, is to characterize the sets $\gamma(A)$ and $\gamma(B)$. This is trivial for finite games, where all behavioral strategies will be included. We shall see that the question is easy for nondegenerate stochastic games. For deterministic continuous games, let us look, for instance, at the setup defined by Hypothesis 2 of §3. A is the set of causal functions from Y^{T-1} into the compact U^{T-1} , Lipschitz continuous with Lipschitz modulus ℓ .

PROPOSITION. The behavioral strategies of $\gamma(A)$ have the following property : let $\tau : U_t \rightarrow \mathbf{R}$ be a function with Lipschitz modulus m , then

$$E^{r^t} \tau = \int_{U_t} \tau(u) d\varphi_t(r^t)(u)$$

is a Lipschitz continuous function of y^t with Lipschitz modulus ℓm .

The proof is elementary. This includes the obvious fact that if the pure strategies are restricted to open loop controls, so are the corresponding behavioral strategies,

since this is the case $\ell = 0$. We conjecture that this is a complete characterization of the set $\gamma(A)$, but this is not sure. We did not investigate this point in detail, since we do not need it in the sequel.

6. Stochastic games. Up to here, we have dealt with games with deterministic dynamics and measurements. The information available to the players is incomplete, but not noisy. We now extend all the previous theory to the case of stochastic games.

A stochastic multistage game is defined as in §2, except that (1) and (2) involve an extra stochastic process $\{w_t\}$, with values in sets \mathcal{W}_t , usually multidimensional. It will always be assumed to be white, with probability distributions W_t known to both players. Thus (1) and (2) are replaced respectively by

$$(17) \quad x_{t+1} = f_t(x_t, u_t, v_t, w_t),$$

$$(18a) \quad y_t = h_t(x_t, w_{t-1}),$$

$$(18b) \quad z_t = k_t(x_t, w_{t-1}).$$

The one time step shift in the argument w of h_t and k_t makes sense, since x_1 being known to both players, the first relevant measurement is y_2, z_2 . Moreover, with that convention, \tilde{h}_t and \tilde{k}_t , as well as r^t and s^t , all depend on u^{t-1}, v^{t-1} , and w^{t-1} , greatly simplifying the sequel. Of course, the hypothesis that w is a white process makes this much more than a pure notational trick. One case where this is not restrictive, and equivalent to the more classical approach, is when w enters in the dynamics and the measurements through distinct independent components.

In this setup, the sequences x, u, v , become stochastic processes, even with pure strategies, and (4) is replaced with a mathematical expectation:

$$(19) \quad G = E \left[\sum_t L_t(x_t, u_t, v_t) + K_{t_1}(x_{t_1}) \right].$$

For simplicity, we shall assume that the sets U_t and V_t do not depend on the current state.

The definitions of the strategies are unchanged. Note that the three sets of hypotheses that have been proposed as alternate setups that ensure satisfaction of the existence hypothesis still stand here. Hypothesis 1 leads to a standard matrix game the entries of which are the expected cost incurred. Hypothesis 2 needs no modification either. Convergence of the trajectories is ensured for each value of w^{T-1} , and thus the expected values converge. Hypothesis 3 must be extended to hold on the $X \times W$ space. It is just a bit cumbersome to state and deal with. A simple case is when $w = (\xi, \eta)$ is made of two independent components, ξ entering in the dynamics, and $h_t(x, w) = \hat{h}_t(x + \eta)$, and likewise for k_t . However, everything becomes much simpler if we assume that for all x_t, u_t, v_t , the transition probability induced by f_t is absolutely continuous with respect to the Lebesgue measure. (The so called *nondegenerate* case). Then, the existence hypothesis and continuity of \tilde{G} for pointwise convergence of the strategies is just a consequence of the Ionescu Tulcea theorem.

Theorem 1 is still valid in this context; its proof is slightly modified as follows.

Proof. (Theorem 1 for stochastic games.) As previously, let Π^{T-1} be the distribution law of the random variable $(u^{T-1}, v^{T-1}, w^{T-1})$ generated by a given pair of mixed strategies (λ, μ) . One has

$$J(\lambda, \mu) = \int_{U^{T-1} \times V^{T-1} \times W^{T-1}} \tilde{G}(u^{T-1}, v^{T-1}, w^{T-1}) d\Pi^{T-1}(u^{T-1}, v^{T-1}, w^{T-1}).$$

Moreover, as previously

$$d\Pi^{T-1}(u^{T-1}, v^{T-1}, w^{T-1}) = d\Pi^{T-2}(u^{T-2}, v^{T-2}, w^{T-2}) d\Pi^c(u_{T-1}, v_{T-1}, w_{T-1}),$$

where Π^c is the conditional law of $(u_{T-1}, v_{T-1}, w_{T-1})$ knowing $(u^{T-2}, v^{T-2}, w^{T-2})$. Using the notation

$$\begin{aligned} p_{T-1} &= \varphi_{T-1}(u^{T-2}, \tilde{h}^{T-1}(u^{T-2}, v^{T-2}, w^{T-2})) \\ q_{T-1} &= \psi_{T-1}(v^{T-2}, \tilde{k}^{T-1}(u^{T-2}, v^{T-2}, w^{T-2})) \end{aligned}$$

for the behaviors, which are, by definition, the conditional laws of u_{T-1}, v_{T-1} for a given r^{T-2} and s^{T-2} , and remembering that w_{T-1} is by hypothesis independent of the past, we derive, still as previously

$$\begin{aligned} d\Pi^{T-1}(u^{T-1}, v^{T-1}, w^{T-1}) = \\ d\Pi^{T-2}(u^{T-2}, v^{T-2}, w^{T-2}) dp_{T-1}(u_{T-1}) dq_{T-1}(v_{T-1}) dR_{T-1}(w_{T-1}). \end{aligned}$$

We may anew iterate the process, to conclude the proof. \square

7. Semicomplete information in finite games. One of the superiorities of behavioral strategies over mixed strategies is that, due to their sequential nature, they lend themselves to dynamic programming. We shall exploit this fact in the case of *finite games with semicomplete information*.

By this we mean that one of the players, say **2**, has full knowledge of the relevant variables of the game at each instant of time. More specifically, we shall assume that at time t , **2** knows x_t, y_t , and also u_{t-1} . This is not beyond the scope of our previous theory, since we may always augment the state with the variables η_t and ξ_t , with $\eta_{t+1} = w_t$, and $\xi_{t+1} = u_t$, so that knowing the full state also yields $y_t = h_t(x_t, \eta_t)$, and $u_{t-1} = \xi_t$. Of course the players are still assumed to have perfect memory, so that they also remember past values of their measurements.

Let ν_t be a probability distribution over $X^t \times V^{t-1}$. We think of ν_t as being player **1**'s conditional probability on (x^t, v^{t-1}) . Assume u^t given, as well as a behavioral strategy ψ^t of the second player. Then, using the dynamics (17), we can propagate ν_t into a probability $\bar{\nu}_{t+1}$ over $X^{t+1} \times V^t$, in the following way. Let $a \in X^{t+1}$ and $b \in V^t$,

$$\bar{\nu}_{t+1}(a, b) = \nu_t(a^t, b^{t-1}) \psi_t[a^t, u^{t-1}, b^{t-1}](b_t) \sum_w \delta(a_{t+1} - f_t(a_t, u_t, b_t, w)) W_t(w).$$

Then, when the measurement y_{t+1} comes in, one may compute the new conditional probability ν_{t+1} on (x^{t+1}, v^t) . We shall give explicit formulas only for the simple example of the next section. Anyhow, this defines a filter of the form

$$(20) \quad \nu_{t+1} = F_t(\nu_t, u^t, y_{t+1}, \psi_t),$$

and also a function

$$(21) \quad \nu_t = N_t(u^{t-1}, y^t, \psi^{t-1}) = N_t(r^t, \psi^{t-1}).$$

By summation over the component subspaces, we can project ν_t on X^t alone, let ρ^t be that law on x^t . We can further project on the component X_t , yielding a law

$$(22) \quad \rho_t = R_t(r^t, \psi^{t-1}).$$

We can now state a theorem of dynamic programming.

THEOREM 2. *Let a stochastic dynamic game be given by (17) to (19), and (φ^*, ψ^*) be a saddle point in behavioral strategies (which exists according to Corollary 1). There exists a sequence of functions V_t from $X^t \times U^{t-1} \times V^{t-1}$ to \mathbf{R} such that for all (x^t, u^{t-1}, v^{t-1}) reached with a non zero probability while playing according to (φ^*, ψ^*) , one has*

$$\begin{aligned}
 (23) \quad & V_t(x^t, u^{t-1}, v^{t-1}) \\
 &= \max_{v \in V_t} \sum_{w \in \mathcal{W}^t} \sum_{u \in U_t} [V_{t+1}(x^t \cdot f_t(x_t, u, v, w_t), u^{t-1} \cdot u, v^{t-1} \cdot v) + L_t(x_t, u, v)] p_t^*(u) W^t(w) \\
 &= \sum_{v \in V_t} \sum_{w \in \mathcal{W}^t} \sum_{u \in U_t} [V_{t+1} + L_t] p_t^*(u) W^t(w) q_t^*(v).
 \end{aligned}$$

(the arguments in V_{t+1} and L_t in the third term are of course the same as in the second) and

$$\begin{aligned}
 (24) \quad & \sum_{x^t, v^{t-1}} V_t(x^t, u^{t-1}, v^{t-1}) \nu_t^*(x^t, v^{t-1}) \\
 &= \min_{u \in U_t} \sum_{w \in \mathcal{W}^t} \sum_{v \in V_t} \sum_{x^t, v^{t-1}} [V_{t+1}(x^t \cdot f_t(x_t, u, v, w_t), u^{t-1} \cdot u, v^{t-1} \cdot v) \\
 &\quad + L_t(x_t, u, v)] \nu_t^*(x^t, v^{t-1}) q_t^*(v) W^t(w) \\
 &= \sum_{u \in U_t} \sum_{w \in \mathcal{W}^t} \sum_{v \in V_t} \sum_{x^t, v^{t-1}} [V_{t+1} + L_t] \nu_t^*(x^t, v^{t-1}) q_t^*(v) W^t(w) p_t^*(u),
 \end{aligned}$$

where p_t^* and q_t^* stand for $\varphi_t^*[r^t]$ and $\psi_t^*[x^t, u^{t-1}, v^{t-1}]$, respectively, ν_t^* for $N_t(r^t, \psi^*)$, and y^t in r^t for $h^t(x^t, w^{t-1})$, and

$$(25) \quad \forall(x, \tau) : (x_\tau, \tau) \in C, \forall(u, v), V_\tau(x^\tau, u^{\tau-1}, v^{\tau-1}) = K_\tau(x_\tau),$$

Conversely, if a sequence of functions V_t together with a pair of behavioral strategies φ^*, ψ^* satisfy equations (23) to (25), these strategies constitute a saddle point of the game, and the value of the game is $V_1(x_1)$.

Proof. The sufficiency part of the claim is a direct adaptation of the theorem in [2] and shall not be repeated in detail. The proof amounts to using (23)–(25) to show that, for an arbitrary sequence v^T , one has $G(\varphi^*, v^T) \leq V_1(x_1)$, and using (24) and (25) to derive the other inequality of the saddle point. This second part uses the fact that when calculating $G(u^T, \psi^*)$, since $\mathbf{2}$ plays ψ^* , $N_t(r^t, \psi^*)$ is actually a conditional probability. This fact is not true, but not needed either, in the first calculation.

Let us now look at necessity.

Notice first that the second equalities in (23) and (24) amount to the fact that q^* has its support contained in the set of v 's that provide the maximum in (23), and likewise for p^* with the minimum in (24).

Let (x^t, u^{t-1}, v^{t-1}) be a state of the game reached with a nonzero probability while playing (φ^*, ψ^*) . For each w^{t-1} , there corresponds a y^{t-1} , and we can describe the game history from there on under the strategies φ^*, ψ^* . Let

$$V_t(x^t, u^{t-1}, v^{t-1}) = \mathbf{E} \left[\sum_{i=t}^{t_1-1} L_i(x_i, u_i, v_i) + K(x_{t_1}) \right].$$

It is clear that V_t thus defined satisfies the second equality of (23), and thus of (24) by summing both sides of (23). Assume now that there exists $\hat{v} \in V_t$ that gives to the second term of (23) a value larger than V_t . Consider the strategy $\hat{\psi}$ that coincides with ψ^* everywhere, except at (x^t, u^{t-1}, v^{t-1}) where it is a dirac distribution at \hat{v} . Let, for simplicity, $L_{t_1} = K(x_{t_1})$, and write

$$\mathbf{E}\tilde{G}(u^{t_1-1}, v^{t_1-1}) = \mathbf{E} \sum_{i=1}^{t-1} L_i + \mathbf{E} \sum_{i=t}^{t_1} L_i.$$

Using the fact that the information algebra is increasing, write the second expectation above as

$$\mathbf{E} \sum_{i=t}^{t_1} L_i = \mathbf{E} \left[\mathbf{E} \left(\sum_{i=t}^{t_1} L_i \mid x^t, u^{t-1}, v^{t-1} \right) \right].$$

The inner expectation is larger for $(\varphi^*, \hat{\psi})$ than for (φ^*, ψ^*) by hypothesis. From all other possible states at time t , this expectation coincides for the two strategy pairs, since ψ^* and $\hat{\psi}$ coincide. However, this particular state is reached by hypothesis with a nonzero probability. Therefore the outer expectation is larger with the strategy $\hat{\psi}$, which contradicts the definition of the saddle point.

We do likewise with φ and (24), noticing as in the sufficiency proof that, when player 2 does play ψ^* , the quantity minimized in (24) actually is the expectation of V_t for player 1. We thus contradict the other inequality of the saddle point. The theorem is proved. \square

Note that, as in [2], this may be viewed as a fixed point theorem: multiply equation (23) by $\nu_t^*(x^t, v^{t-1})$ on both sides, and sum over all (x^t, v^{t-1}) . Then (23) and (24) together express the fact that φ_t^*, ψ_t^* provide a saddle point (over a product of simplices for ψ_t^*) of the matrix made up of the blocks $[V_{t+1} + L_t]$ weighted by $\nu_t^* = N_t(r^t, \psi^{t-1})$. So the problem is to find a ψ^* that gives rise to a ν^* for which this ψ^* is an argument of this sequence of saddle points.

In itself, this theorem is of little use. We shall see in the next section a case where it simplifies to the point where it can be used to compute the saddle point of the game. At this time, we show a theoretical consequence of interest.

COROLLARY 2. *Let $\rho_t^* = R_t(r^t, \psi^*)$. The game admits a saddle point in behavioral strategies of the form $\varphi_t^*[r^t] = \hat{\varphi}_t[\rho_t^*]$, $\psi_t^*[x^t, u^{t-1}, v^{t-1}] = \hat{\psi}_t[x_t, \rho_t^*]$. We can define a filter $\rho_{t+1} = g_t(\rho_t, u_t, y_{t+1}, \psi_t)$, and there exists a sequence of functions $V_t(x_t, \rho_t)$ such that for all (x_t, ρ_t) that are reached with a nonzero probability while playing optimally,*

$$\begin{aligned} (26) \quad & V_t(x_t, \rho_t) \\ &= \max_{v \in V_t} \sum_{w \in \mathcal{W}_t} \sum_{u \in U_t} [V_{t+1}(f_t(x_t, u, v, w), g_t(\rho_t, u, y_{t+1}, \hat{q}_t)) + L_t(x_t, u, v)] \hat{p}_t(u) W_t(w) \\ &= \sum_{v \in V_t} \sum_{w \in \mathcal{W}_t} \sum_{u \in U_t} [V_{t+1} + L_t] \hat{p}_t(u) W_t(w) \hat{q}_t(v) \end{aligned}$$

and

$$(27) \quad \sum_{x_t \in X_t} V_t(x_t, \rho_t) \rho_t(x) =$$

$$\begin{aligned} \min_{u \in U_t} \sum_{x_t} \sum_{w \in W_t} \sum_{v \in V_t} [V_{t+1}(f_t(x_t, u, v, w), g_t(\rho_t, u, y_{t+1}, \hat{q}_t)) + L_t(x_t, u, v)] \hat{q}_t(v) W_t(w) \rho_t(x_t) \\ = \sum_{u \in U_t} \sum_{x_t} \sum_{w \in W_t} \sum_{v \in V_t} [V_{t+1} + L_t] \hat{q}_t(v) W_t(w) \rho_t(x_t) \hat{p}(u). \end{aligned}$$

where \hat{p}_t stands for $\hat{\varphi}_t[\rho_t]$, \hat{q}_t for $\hat{\psi}_t[x_t, \rho_t]$, and y_{t+1} for $h_{t+1}(f_t(x_t, u, v, w), w)$, and

$$(28) \quad \forall(x_\tau, \tau) \in C, \forall \rho, V_\tau(x_\tau, \rho) = K_\tau(x_\tau).$$

Conversely, if a sequence of functions $\hat{\varphi}_t$, $\hat{\psi}_t$, and V_t satisfy these equations, they provide a saddle point of the game.

Proof. The fact that with strategies of this form there exists such a filter and the sufficiency part of the proof is exactly the main theorem of [2]. Let us look at the necessity.

Notice that, for each value of (x^t, v^{t-1}) in $X^t \times V^{t-1}$, we may multiply all terms of equation (23) by $\nu_t^*(x^t, v^{t-1})$, which is nonnegative, and sum over all such terms still preserving the inequalities. Conversely, writing the resulting summed inequality (implicit in the max operation) is *equivalent* to the separate inequalities, provided it is specified that q^* is allowed to depend on (x^t, v^{t-1}) . (Therefore the combined maximizing variable ranges over a product of simplices.)

Therefore, the mean value $\bar{V}(r^t) = \sum V_t \nu_t^*$ appears as the saddle point of the kernel $\sum_w \sum_{(x^t, v^{t-1})} [V_{t+1} + L_t] \nu_t^*(x^t, v^{t-1}) W^t(w)$, over a simplex for the minimizing variable, and a product of simplices for the maximizing variable. Now look at this definition for $t = T - 1$. Then, $V_{t+1} = K_T(x_T)$, and the kernel depends on past values of the various variables only through ν_{T-1}^* . Moreover, since the variables x^{T-2} , v^{T-2} , and w^{T-2} do not appear in V_T and L_{T-1} , we may first sum over these variables, ending up with the kernel $\sum_w \sum_x [V_T + L_{T-1}] \rho_{T-1}(x) W_{T-1}(w)$. Therefore, the value of the saddle point depends only on ρ_{T-1} . And using the sufficiency argument, we can replace φ_{T-1}^* and ψ_{T-1}^* by strategies of the form proposed for $\hat{\varphi}$ and $\hat{\psi}$. So V_{T-1} also depends only on x_{T-1} and ρ_{T-1} .

Finally, using the propagation of ρ_t as stated in the theorem, we can iterate this process for time $T - 2$, and so on down to 1. This proves the theorem. \square

This theorem shows that (x_t, ρ_t) constitutes a *sufficient statistic* of constant dimension for the decision problem at hand. The dynamic programming algorithm that one would like to derive from this theory remains quite cumbersome for two reasons. On the one hand, one must work in a space with a *continuous* component, while the game is discrete, and even finite. On the other hand, as was pointed out in [2], each step involves the solution of a difficult fixed point problem, since $(\hat{\varphi}_t, \hat{\psi}_t)$ must be the saddle point of a kernel that itself depends on $\hat{\psi}_t$ through its appearance in g_t . The above theory proves that this fixed point exists, a result which could not be obtained through the classical topological techniques, since the dependence of the kernel on $\hat{\psi}$ is not continuous. But computing it may remain a formidable task.

So we turn now to an example where one may prefer to stick with the full behavioral strategies.

8. Rabbit and Hunter game. This game is an extension of the example of §4. A hunter tries to shoot a rabbit that moves in a finite space made of N positions, assumed for this example to lie on a straight line. The game is specified by six integers:

- T , the horizon of the game. As usual, $\mathbf{T} = \{1, \dots, T\}$.
- N , the number of possible positions of rabbit. The game space is therefore $\mathbf{N} = \{1, \dots, N\}$.

- a , the amplitude of rabbit's jumps, (or a^- , a^+ , the left and right amplitudes).
- b , the number of bullets available to hunter,
- c , the capture radius (or lethal radius) of a bullet,
- d , the delay or time taken by the bullets to fly from hunter to rabbit.

The state x_t is composed of the following scalar variables:

- $y_t \in \mathbf{N}$, the position of rabbit.
- $w_t^k \in \mathbf{N} \cup \{0\}$, $k = 1, \dots, d-1$, the position the bullet shot k time steps earlier is flying to.
- $z_t \in \mathbf{N} \cup \{0\}$, the position where the bullet shot d time steps earlier is arriving. (It will be convenient to use this notation rather than w^d .)
- κ_t , the counter of expended bullets.

The players controls are

- $u_t \in U_t = [y_t - a, y_t + a] \cap \mathbf{N}$, (or $U_t = [y_t - a^-, y_t + a^+] \cap \mathbf{N}$), the next position of rabbit.
- $v_t \in V_t$, the position hunter aims at. $v_t = 0$ means that he does not shoot, since 0 is not a possible position of rabbit. To take into account the budget constraint, we set

$$V_t = \begin{cases} \mathbf{N} \cup \{0\} & \text{if } \kappa_t < b, \\ \{0\} & \text{if } \kappa_t = b. \end{cases}$$

The dynamics are

$$\begin{aligned} y_{t+1} &= u_t, & y_1 & \text{ given,} \\ w_{t+1}^1 &= v_t, & w_1^1 &= 0, \\ w_{t+1}^{k+1} &= w_t^k & w_1^k &= 0, \\ z_{t+1} &= w_t^{d-1} & z_1 &= 0, \\ \kappa_{t+1} &= \kappa_t + 1 - \delta(v_t), & \kappa_1 &= 0. \end{aligned}$$

Having included the budget constraint into V_t , we may take for the capture set $\{|z_t - y_t| \leq c\}$, pretending that the game goes on, even if hunter cannot shoot anymore.

In fact, to simplify the calculations below, we shall from now on assume that

$$b \leq N,$$

so that hunter shoots at all time steps, and we may ignore κ .

We shall need the following notations:

$$C(w) = \{u \mid |u - w^{d-1}| \leq c\}$$

and

$$\chi_w(u) = \begin{cases} 1 & \text{if } u \in C(w), \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, with a slight abuse of notations, we shall write either $u \in C(w)$ or $w \in C(u)$, meaning $(u, w) \in C$.

Introduce finally the following shift operator operating on the vector w :

$$\sigma(v) \cdot w = \begin{pmatrix} v \\ w^1 \\ \vdots \\ w^{d-2} \end{pmatrix}.$$

In the sequel, let

$$\begin{aligned} p_t^* &= \varphi_t^*[y^t], \\ q_t^* &= \psi_t^*[x^t]. \end{aligned}$$

The filter must take into account that the rabbit computes ρ_{t+1} only if it is alive, which gives extra information on the past. We get (see [3])

$$(29) \quad \rho_{t+1}(w) = \sum_{\omega \notin C(y_{t+1})} \delta(w - \sigma(w^1) \cdot \omega) q_t^*(w^1) \rho_t(\omega) \left[\sum_{\omega \notin C(y_{t+1})} \rho_t(\omega) \right]^{-1}.$$

The dynamic programming equations (23) and (25) now read

$$V_t(y^t, w^t, z^t) = \max_{v \in V_t} \sum_{u \in U_t} V_{t+1}(y^t \cdot u, w^t \cdot \sigma(v) \cdot w, z^t \cdot w^{d-1}) p_t^*(u)$$

if $y_t \notin C(z_t)$, i.e., $(y_t, z_t) \notin C$,

$$V_t(y^t, w^t, z^t) = 1 \quad \text{if } (y_t, z_t) \in C.$$

We may simplify this expression, and more importantly reduce the size of the space to be scanned, in the following manner. Introduce a function $W_t(y^t, w_t)$, which will be related to the function V_t according to the following definition:

$$(30) \quad V_t(y^t, w^t, z^t) = \begin{cases} W_t(y^t, w_t) & \text{if } (y_t, z_t) \notin C, \\ 1 & \text{if } (y_t, z_t) \in C. \end{cases}$$

This function satisfies the following dynamic programming equations, which show that it only depends on the indicated variables, exactly in the same way as we proved Corollary 2 above:

$$W_t(y^t, w_t) = \max_{v \in V_t} \sum_{u \in U_t} [\chi_{w_t}(u) + (1 - \chi_{w_t}(u)) W_{t+1}(y^t \cdot u, \sigma(v) \cdot w_t)] p_t^*(u),$$

or equivalently

$$(31) \quad W_t(y^t, w_t) = \max_{v \in V_t} \left[\sum_{u \in C(w_t)} p_t^*(u) + \sum_{u \notin C(w_t)} W_{t+1}(y^t \cdot u, \sigma(v) \cdot w_t) p_t^*(u) \right].$$

The second dynamic programming equation, (24), becomes now

$$(32) \quad \sum_w W_t(y^t, w) \rho_t(w) = \min_{u \in U_t} \sum_w [\chi_w(u) + (1 - \chi_w(u)) \sum_v W_{t+1}(y^t \cdot u, \sigma(v) \cdot w) q_t^*(v)] \rho_t(w).$$

Equations (31) and (32) may be used as the basis for a numerical algorithm, provided that the horizon T be short enough, (and a and N small enough) so that the state space remain of a tractable size. The algorithm again involves a fixed point search: for a given set of ρ_t 's at each point of the space, compute $\hat{\varphi}$ and $\hat{\psi}$ rearwards in time, then solve for the fixed point $N_t(y^t, \hat{\psi}) = \rho_t$. This can be done, for instance, via a successive approximation scheme, using subrelaxation as necessary. No convergence proof is available at this time, however.

Notice finally that, as in [3], the mean value

$$\sum_w W_t(y^t, w) \rho_t(w) = \bar{W}_t(y^t)$$

can be computed in a faster way, avoiding the separate computations according to the values of w . Taking ρ_{t+1} in (29), and for a fixed $y_{t+1} = u$, we have

$$\sum_w W_{t+1}(y^t \cdot u, \omega) \rho_{t+1}(\omega) = [1 - \rho_t(C_t(u))]^{-1} \sum_w \sum_{w \notin C(u)} \sum_v W_{t+1}(y^t \cdot u, \omega) \delta(\omega - \sigma(v) \cdot w) q_t^*(v) \rho_t(w).$$

Take the summation in ω inside the other two, and use the fact that then, the δ selects the only value $\omega = \sigma(v) \cdot w$, to get

$$\bar{W}_{t+1}(y^t \cdot u) = [1 - \rho_t(C_t(u))]^{-1} \sum_{w \notin C(u)} \sum_v W_{t+1}(y^t \cdot u, \sigma(v) \cdot w) q_t^*(v) \rho_t(w),$$

or equivalently

$$[1 - \rho_t(C_t(u))] \bar{W}_{t+1}(y^t \cdot u) = \sum_w (1 - \chi_w(u)) \sum_v W_{t+1}(y^t \cdot u, \sigma(v) \cdot w) q_t^*(v) \rho_t(w).$$

We recognize the right-hand side here as being the second term in the right-hand side of (32) above. Substituting into it we get a recurrent equation for \bar{W}_t :

$$\bar{W}_t(y^t) = \min_u [\rho_t(C(u)) + [1 - \rho_t(C_t(u))] \bar{W}_{t+1}(y^t \cdot u)].$$

Acknowledgment. This work owes much to fruitful discussions with Tamer Başar, of the University of Illinois, then on leave at INRIA Sophia Antipolis, who, among other things, suggested the example of §4 which led to the understanding of the nature of mixed versus behavioral strategies in dynamic games.

REFERENCES

- [1] R. J. AUMANN, *Mixed and behavior strategies in infinite extensive games*, Advances in Game Theory, Ann. Math. Stud. 52, (1964), pp. 627–650.
- [2] P. BERNHARD AND A. L. COLOMB, *Saddle point condition for a class of stochastic dynamical games with imperfect information*, IEEE Trans. Automat. Control, AC 10-23 (1987), pp. 98–101.
- [3] P. BERNHARD, *Computation of equilibrium points of delayed partial information games*, IEEE Conf. Decision and Control, Los Angeles, CA, 1987.
- [4] I. EKELAND, *La théorie des jeux et ses applications à l'économie mathématique*, P.U.F., Paris, 1974.
- [5] H. W. KUHN, *Extensive games and the problem of information*, Ann. Math. Stud. 28 (1953), pp. 193–216.
- [6] J. LÉVINE, *Incomplete information and optimality, a dynamic programming approach to necessary and sufficient conditions*, internal report A/93, Centre d'Automatique et Informatique of Ecole Nationale Supérieure des Mines de Paris, Fontainebleau, France, 1980.
- [7] J. NEVEU, *Bases mathématiques du calcul des probabilités*, Masson, Paris, 1964.
- [8] A. S. NOWAK, *Existence of optimal strategies in zero sum nonstationary stochastic games with lack of information on both sides*, SIAM J. Control Optim., 27 (1989), pp. 289–295.
- [9] M. SCHÄL, *Stochastic nonstationary two person zero sum games*, Z. Angew. Math. Mech., 61 (1981), pp. 352–353.
- [10] M. SION, *On general minimax theorems*, Pacific J. Math., 8 (1958), pp. 171–176.
- [11] J. VON NEUMANN AND O. MORGENSTERN, *Theory of Games and Economic Behaviour*, 2nd ed., Princeton, 1947.

ON THE MODELLING AND EXACT CONTROLLABILITY OF NETWORKS OF VIBRATING STRINGS*

E. J. P. GEORG SCHMIDT†

Abstract. Using Hamilton's principle a nonlinear system of partial differential equations describing the dynamics of a network of vibrating strings is derived. An equilibrium is linearized to obtain a linear system of "wave equations." The equations are complemented by "coupling conditions" at the "multiple nodes" where several elements meet and by "boundary conditions" at the "simple nodes." Existence of solutions to the linear system is proved. Finally, multiplier techniques are used to prove exact controllability for certain specific networks.

Key words. boundary control, hyperbolic systems, vibrating networks

AMS(MOS) subject classifications. 36B37, 36L55, 49E15, 93C20, 73K03

1. Introduction. Our purpose is to describe the dynamics, and then to control exactly connected networks of vibrating strings in configurations such as those in Fig. 1.

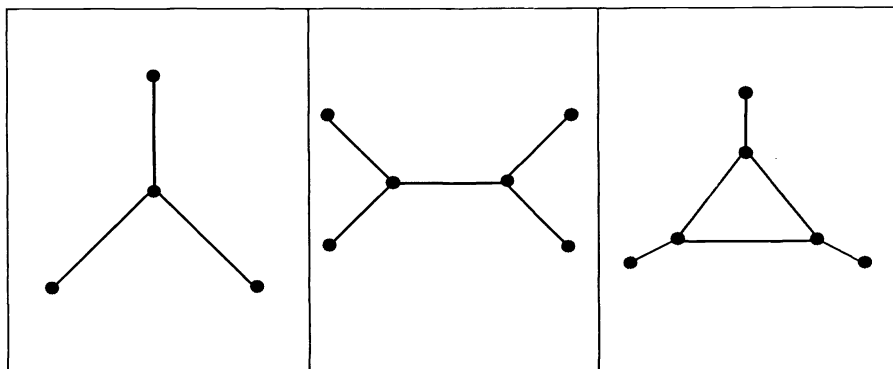


FIG. 1

Our interest in this problem was stimulated by Chen, Delfour, Krall, and Payre [3], who dealt with the stabilization of "serial" beam configurations. We wanted to treat more complicated networks and have done so, in the technically somewhat simpler context of vibrating strings. We have focussed on exact controllability; results on stabilizability could surely be obtained using similar methods. A preliminary version

*Received by the editors December 11, 1989; accepted for publication (in revised form) November 30, 1990. This work was supported by Natural Sciences and Engineering Research Council grant A727, as well as by the Deutsche Forschungsgemeinschaft (Schwerpunktsprogramm: Anwendungsbezogene Optimierung und Steuerung).

†Department of Mathematics and Statistics, McGill University, 805 Sherbrooke Street West, Montreal, Quebec, Canada, H3A 2K6.

of this paper has appeared in a technical report [9]. Networks of beams are now the subject of a joint work with G. Leugering; some results on beam networks have appeared in [5].

The endpoints of the component strings will be referred to as *nodes*; the nodes where two or more strings meet will be called *multiple*, while the nodes corresponding to only one string will be called *simple*. In this paper the simple nodes will be fixed or subject to control, while the multiple nodes will be allowed to oscillate freely. We consider small perturbations of equilibrium configurations and linearize a certain system of nonlinear equations to obtain a hyperbolic system of linear equations that govern the evolution of the displacements from the equilibrium. It is that linear system that we shall deal with in detail. Here we deal with planar vibrations of planar networks—the generalisation to three dimensions involves no essential difficulties.

2. Modelling networks of vibrating strings. We begin by considering planar vibrations of a single elastic string segment. If the “natural” length of that string is ℓ , it is reasonable to consider the string as parametrised by the rest arc length σ with σ in $[0, \ell]$. The position at time t of that point corresponding to the rest position σ is to be denoted by the vector $\mathbf{R}(\sigma, t)$. If ρ is the density of the string, and if the string is assumed to satisfy Hooke’s law with constant h , the kinetic and potential energies of the string are given by

$$\frac{1}{2} \int_0^\ell \rho |\mathbf{R}_t(\sigma, t)|^2 d\sigma \quad \text{and} \quad \frac{1}{2} \int_0^\ell h [|\mathbf{R}_\sigma(\sigma, t)| - 1]^2 d\sigma,$$

respectively; here the subscripts indicate partial differentiation with respect to the indicated variables, and $|\cdot|$ denotes the Euclidean norm. The total energy is then

$$(1) \quad E(t) = \frac{1}{2} \int_0^\ell [\rho |\mathbf{R}_t(\sigma, t)|^2 + h [|\mathbf{R}_\sigma(\sigma, t)| - 1]^2] d\sigma.$$

Applying Hamilton’s principle to the Lagrangian

$$\mathcal{L}(\mathbf{R}(\sigma, t)) = \frac{1}{2} \int_0^T \int_0^\ell [\rho |\mathbf{R}_t(\sigma, t)|^2 - h [|\mathbf{R}_\sigma(\sigma, t)| - 1]^2] d\sigma dt,$$

we obtain the nonlinear system of partial differential equations

$$(2) \quad \rho \mathbf{R}_{tt} = h [(|\mathbf{R}_\sigma| - 1) \mathbf{R}_\sigma / |\mathbf{R}_\sigma|]_\sigma.$$

This needs to be complemented by suitable boundary conditions at $\sigma = 0$ and $\sigma = \ell$, as well as by initial conditions on $\mathbf{R}(\sigma, 0)$ and $\mathbf{R}_t(\sigma, 0)$. For time independent Dirichlet or Neumann boundary conditions, it is easy to check that energy is conserved when the equations are satisfied.

Remarks.

- (a) The system is of second order in both t and σ .
- (b) The assumption that Hooke’s law holds is certainly only reasonable for a limited range of extension (corresponding to $|\mathbf{R}_\sigma| > 1$) or contraction (corresponding to $|\mathbf{R}_\sigma| < 1$). Hooke’s law can be replaced by using other potential energy functions of the form $U(|\mathbf{R}_\sigma| - 1)$; this leads to a variety of interesting nonlinear systems.
- (c) For related nonlinear string models derived in alternate ways, see Carrier [2] and Antman [1].

A network consists of many string segments, each of which is parametrized and described as above. These strings may have different physical characteristics and therefore be associated with different constants ℓ_i , ρ_i , and h_i . Let the spatial displacement of the i th string be $\mathbf{R}^i(\sigma, t)$. Then the total energy of a network consisting of n strings is

$$(3) \quad E(t) = \frac{1}{2} \sum_{i=1}^n \int_0^{\ell_i} [\rho_i |\mathbf{R}_t^i(\sigma, t)|^2 + h_i [|\mathbf{R}_\sigma^i(\sigma, t)| - 1]^2] d\sigma,$$

and the corresponding Lagrangian is

$$(4) \quad \mathcal{L}(\{\mathbf{R}^i(\sigma, t)\}) = \frac{1}{2} \sum_{i=1}^n \int_0^T \int_0^{\ell_i} [\rho_i |\mathbf{R}_t^i(\sigma, t)|^2 - h_i [|\mathbf{R}_\sigma^i(\sigma, t)| - 1]^2] d\sigma.$$

We introduce the notation $\mathcal{E}(\mathbf{N}) = \{i : \text{the } i\text{th string meets } \mathbf{N}\}$, where $\mathbf{R}^i(\mathbf{N}, t)$ is equal to either $\mathbf{R}^i(\ell_i, t)$ or $\mathbf{R}^i(0, t)$ depending on the parameter value 0 or ℓ_i corresponding to \mathbf{N} . For the networks under consideration, the displacements $\mathbf{R}^i(\sigma, t)$ should clearly satisfy the following “geometric node condition” at each multiple node \mathbf{N} :

$$(5) \quad \text{the } \mathbf{R}^i(\mathbf{N}, t) \text{ are equal for all } i \in \mathcal{E}(\mathbf{N}).$$

Applying Hamilton’s principle to the Lagrangian (4) with the “variations” satisfying (5), we obtain the partial differential equations

$$(6) \quad \rho_i \mathbf{R}_{tt}^i = h_i [(|\mathbf{R}_\sigma^i| - 1) \mathbf{R}_\sigma^i / |\mathbf{R}_\sigma^i|]_\sigma, \quad i = 1 \text{ to } n,$$

as well as the following “dynamic” condition at multiple nodes \mathbf{N} :

$$(7) \quad \sum_{i \in \mathcal{E}(\mathbf{N})} \epsilon_i(\mathbf{N}) h_i [(|\mathbf{R}_\sigma^i(\mathbf{N}, t)| - 1) \mathbf{R}_\sigma^i(\mathbf{N}, t) / |\mathbf{R}_\sigma^i(\mathbf{N}, t)|] = 0,$$

where $\epsilon_i(\mathbf{N}) = -1$ or $+1$ depending on whether \mathbf{N} equals $\mathbf{R}^i(0, t)$ or $\mathbf{R}^i(\ell_i, t)$ (in which case $\epsilon_i(\mathbf{N}) \mathbf{R}_\sigma^i(\mathbf{N}, t)$ becomes the “outward pointing” derivative of \mathbf{R}).

Furthermore, the simple nodes can either be held fixed or controlled; this corresponds to imposing a Dirichlet condition at each simple node with Dirichlet data that is either fixed in time, or time dependent and subject to choice. Initial “states” $(\mathbf{R}^i(\sigma, 0), \mathbf{R}_t^i(\sigma, 0))$ also must be prescribed for each string segment. If the Dirichlet data are all taken as time independent, then it is easy to verify, at least formally, that the total energy of the network is conserved. Letting $E^i(t)$ denote the energy of the i th string we check that

$$\frac{d}{dt} E^i(t) = h_i [\mathbf{R}_t^i(\sigma, t) \cdot (|\mathbf{R}_\sigma^i(\sigma, t)| - 1) \mathbf{R}_\sigma^i(\sigma, t) / |\mathbf{R}_\sigma^i(\sigma, t)|]_{\sigma=0}^{\sigma=\ell_i}.$$

Now note that $\mathbf{R}_t^i(\sigma, t) = 0$ at the simple nodes, and add the contributions at each multiple node; using (5) and (7) we get that the sum of all these endpoint terms is zero.

For an equilibrium configuration of the network, each $\mathbf{R}^i(\sigma, t)$ will have the time independent form

$$(8) \quad \mathbf{R}^i(\sigma, t) = \mathbf{R}_0^i + \sigma s_i \mathbf{v}^i,$$

where \mathbf{v}^i is the unit direction pointing along the i th string segment and s_i is a constant measuring the amount of uniform stretching of that segment. We assume that $s_i > 1$, so that the string is under tension. Time independent boundary conditions, as well as the compatibility conditions at multiple nodes, need to be satisfied. We wish to linearize about such an equilibrium configuration.

To carry out the linearization, it is enough to consider a single string segment and assume that

$$(9) \quad \mathbf{R}(\sigma, t) = \mathbf{R}_0 + \sigma s \mathbf{v} + \mathbf{r},$$

where $\mathbf{r}(\sigma, \mathbf{t}) = (\mathbf{x}(\sigma, \mathbf{t}), \mathbf{y}(\sigma, \mathbf{t}))$. Then $\mathbf{R}_{tt}(\sigma, t) = \mathbf{r}_{tt}(\sigma, t) = (x_{tt}(\sigma, t), y_{tt}(\sigma, t))$. The right-hand side of (2) can either be linearized directly using a Taylor expansion in x and y of

$$|\mathbf{r}_\sigma| = [\mathbf{r}_\sigma \cdot \mathbf{r}_\sigma]^{1/2} = [s^2 + 2s\mathbf{v} \cdot (x, y) + x^2 + y^2]^{1/2}$$

about $(0,0)$, or, alternatively, we can similarly expand the potential energy up to quadratic terms in x_σ and y_σ to obtain

$$h[|\mathbf{R}_\sigma| - 1]^2 = \frac{k}{2}[(s-1)^2 + 2(s-1)\mathbf{v} \cdot (x_\sigma, y_\sigma) + (1 - \frac{1}{s})(x_\sigma^2 + y_\sigma^2) + \frac{1}{s}(\mathbf{v} \cdot (x_\sigma, y_\sigma))^2] + \dots$$

Either way, we obtain the system

$$\begin{aligned} \rho x_{tt} &= \frac{k}{s}(s-1+a^2)x_{\sigma\sigma} + \frac{k}{s}aby_{\sigma\sigma'}, \\ \rho y_{tt} &= \frac{k}{s}abx_{\sigma\sigma} + \frac{k}{s}(s-1+b^2)y_{\sigma\sigma'}, \end{aligned}$$

where $\mathbf{v} = (a, b)$. This linear system can be rewritten as

$$(10) \quad \rho \mathbf{r}_{tt} = H \mathbf{r}_{\sigma\sigma} \quad \text{with} \quad H = \begin{bmatrix} s-1+a^2 & ab \\ ab & s-1+b^2 \end{bmatrix}.$$

We note that H is a symmetric matrix with eigenvalues h and $h(1-s^{-1})$ (both positive since $s > 1$) corresponding respectively to eigenvectors \mathbf{v} and \mathbf{v}^\perp (a vector perpendicular to \mathbf{v}). This means that for a particular vibrating string segment the vibration can be uncoupled into transverse and longitudinal vibrations. For the potential energy of such a linearized string segment we keep only the quadratic term and, hence, the total energy of that segment is

$$(11) \quad E(t) = \frac{1}{2} \int_0^\ell [\rho |\mathbf{r}_t(\sigma, t)|^2 + H \mathbf{r}_\sigma(\sigma, t) \cdot \mathbf{r}_\sigma(\sigma, t)] d\sigma.$$

The linear equations governing the network near equilibrium now become

$$(12) \quad \rho_i \mathbf{r}_{tt}^i = H^i \mathbf{r}_{\sigma\sigma}^i \quad \text{for } i = 1 \text{ to } n.$$

It is notationally convenient to suppose that the simple nodes are $\mathbf{r}^i(\ell_i, t)$ with $i = 1$ to m , and that controls are applied at these points for $i = 1$ to p . The *boundary conditions* at the simple nodes then take the form

$$(13) \quad \mathbf{r}^i(\ell_i, t) = \begin{cases} \mathbf{u}^i(t) & \text{for } i = 1 \text{ to } p, \\ 0 & \text{for } i = p+1 \text{ to } m. \end{cases}$$

At the multiple nodes we require two *coupling conditions*: (5) rewritten as

$$(14) \quad \text{the } \mathbf{r}^i(\mathbf{N}, t) \text{ are equal for all } i \in \mathcal{E}(\mathbf{N});$$

and the linearization of (7), namely

$$(15) \quad \sum_{i \in \mathcal{E}(\mathbf{N})} \epsilon_i(\mathbf{N}) H^i \mathbf{r}_\sigma^i(\mathbf{N}, t) = 0.$$

The *initial conditions* accompanying (12) are

$$(16) \quad (\mathbf{r}^i(\sigma, 0), \mathbf{r}_t^i(\sigma, 0)) = (\mathbf{r}^{i,0}(\sigma), \mathbf{r}^{i,1}(\sigma)) \quad \text{for } i = 1 \text{ to } n.$$

We note that under the above conditions, with $\mathbf{u}^i(t) = \mathbf{c}^i$, where \mathbf{c}^i are given constant vectors, energy is conserved much as before.

3. Existence of solutions to the linearized system. We need an existence theory for the linear system described by (12)–(16). We begin by introducing the spaces in which solutions should lie.

Let $L_2(0, \ell_i)$ and $W^1(0, \ell_i)$ be the familiar Lebesgue and Sobolev spaces, with the understanding that in our context they consist of 2-vector valued functions. We then introduce the space \mathcal{H}_E as the set of all data $\{(\mathbf{r}^{i,0}(\sigma), \mathbf{r}^{i,1}(\sigma))\}$ (where $\{\cdots\}$ will denote n-tuples $\{\cdots\}_{i=1}^n$) belonging to the product of the n data spaces $W^1(0, \ell^i) \times L_2(0, \ell_i)$ for the separate segments satisfying the first coupling condition (14) at all multiple nodes. On this space we can define the “energy scalar product”

$$(17) \quad \begin{aligned} & \langle \{(\mathbf{r}^{i,0}(\sigma), \mathbf{r}^{i,1}(\sigma))\}, \{(\bar{\mathbf{r}}^{i,0}(\sigma), \bar{\mathbf{r}}^{i,1}(\sigma))\} \rangle_E \\ &= \frac{1}{2} \sum_{i=1}^n \int_0^{\ell_i} [\rho_i \mathbf{r}_t^i(\sigma, t) \cdot \bar{\mathbf{r}}_t^i(\sigma, t) + H^i \mathbf{r}_\sigma^i(\sigma, t) \cdot \bar{\mathbf{r}}_\sigma^i(\sigma, t)] d\sigma. \end{aligned}$$

This fails to be an inner product on \mathcal{H}_E because the associated norm vanishes for the “constant” data $\{(\mathbf{c}, \mathbf{0})\}$, which corresponds physically to a uniform displacement of the whole network by \mathbf{c} . It is convenient to identify $\{(\mathbf{r}^{i,0}(\sigma), \mathbf{r}^{i,1}(\sigma))\}$ with $\{\mathbf{r}^{i,0}(\sigma)\} \times \{\mathbf{r}^{i,1}(\sigma)\}$. We require the second factor in this product to belong to the space \mathcal{H}_1 defined as the product space of the spaces $L_2(0, \ell^i)$ with inner product

$$\langle \{\mathbf{r}^i(\sigma)\}, \{\bar{\mathbf{r}}^i(\sigma)\} \rangle_1 = \frac{1}{2} \sum_{i=1}^n \int_0^{\ell_i} [\rho_i \mathbf{r}^i(\sigma) \cdot \bar{\mathbf{r}}^i(\sigma)] d\sigma.$$

The first factor is required to belong to a space that is consistent with the boundary condition (13), as well as the coupling condition (14). Let \mathcal{H}_0 denote the product space of the spaces $W^1(0, \ell_i)$ consisting of functions satisfying condition (14) at multiple nodes. If $p < m$ let \mathcal{H}_0^p be the subspace of \mathcal{H}_0 consisting of data satisfying $\mathbf{r}^i(\ell_i) = \mathbf{0}$ for $i = p+1$ to m) (i.e., at the fixed simple nodes). Otherwise, when $p = m$ and none of the simple nodes is fixed, we define \mathcal{H}_0^m to be the quotient space of \mathcal{H}_0 and the subspace of constant data $\{\mathbf{c}\}$; for notational simplicity we denote elements of this space by representatives of the corresponding equivalence class. In either case, we can introduce as the inner product

$$\langle \{\mathbf{r}^i(\sigma)\}, \{\bar{\mathbf{r}}^i(\sigma)\} \rangle_0 = \frac{1}{2} \sum_{i=1}^n \int_0^{\ell_i} [H^i \mathbf{r}_\sigma^i(\sigma) \cdot \bar{\mathbf{r}}_\sigma^i(\sigma)] d\sigma;$$

this inner product yields a norm equivalent to the standard inner product norm by a variant of the Poincaré lemma, which can be proved in the usual way (for example, by adapting the arguments of Necas [7, pp. 18–21]). Then (17) defines the “energy inner product” on the space $\mathcal{H}_E^p = \mathcal{H}_0^p \times \mathcal{H}_1$.

We pursue an analogy with the wave equation in a domain where the simple nodes of a network correspond to the boundary of the domain. We expect solutions to homogeneous boundary value problems to be given in terms of a unitary semigroup acting on initial data. We can rewrite (12) as

$$\frac{d}{dt} \begin{bmatrix} \mathbf{r}^i(\sigma, t) \\ \mathbf{r}_t^i(\sigma, t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \frac{1}{\rho_i} H^i \partial_\sigma^2 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{r}^i(\sigma, t) \\ \mathbf{r}_t^i(\sigma, t) \end{bmatrix}, \quad \text{for } i = 1 \text{ to } n,$$

where ∂_σ is the operation of partial differentiation with respect to σ . Alternatively, this system can be written as

$$(18) \quad \frac{d}{dt} \begin{bmatrix} \{\mathbf{r}^i(\sigma, t)\} \\ \{\mathbf{r}_t^i(\sigma, t)\} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \mathcal{A} & 0 \end{bmatrix} \begin{bmatrix} \{\mathbf{r}^i(\sigma, t)\} \\ \{\mathbf{r}_t^i(\sigma, t)\} \end{bmatrix},$$

where \mathcal{A} is the linear map defined by

$$\mathcal{A}(\{\mathbf{r}^i(\sigma)\}) = \left\{ \frac{1}{\rho_i} H^i \partial_\sigma^2 \mathbf{r}^i(\sigma) \right\}$$

on a suitable domain to be specified below, with the understanding that 0 and 1 are to be understood as null and identity operators. This suggests the form that the skew adjoint generator of the appropriate semigroup should have. We now need to study the operator \mathcal{A} , which plays a role in networks analogous to that of the Laplacian for Euclidean domains.

Given $\{\mathbf{f}^i(\sigma)\}$ in \mathcal{H}_1 and constants \mathbf{c}^1 to \mathbf{c}^m , we say that $\{\mathbf{r}^i(\sigma)\}$ is a weak solution of the problem

$$(19) \quad \mathcal{A}\{\mathbf{r}^i(\sigma)\} = \{\mathbf{f}^i(\sigma)\}, \quad \mathbf{r}^i(\ell_i) = \mathbf{c}^i \text{ for } i = 1 \text{ to } m$$

subject to the node conditions (14) and (15), if and only if $\{\mathbf{r}^i(\sigma)\}$ belongs to \mathcal{H}_0 , satisfies the boundary conditions $\mathbf{r}^i(\ell_i) = \mathbf{c}^i$ for $i = 1$ to m as well as the weak identity

$$\langle \{\mathbf{r}^i(\sigma)\}, \{\phi^i(\sigma)\} \rangle_0 = -\langle \{\mathbf{f}^i(\sigma)\}, \{\phi^i(\sigma)\} \rangle_1, \quad \text{for all } \{\phi^i(\sigma)\} \text{ in } \mathcal{H}_0^0.$$

The assertions of the following theorem are proved in exactly the same way as the corresponding results on the Dirichlet problem for the Laplacian in a bounded domain; the principal tools are the Riesz representation theorem as well as Rellich’s compactness theorem.

THEOREM 3.1. *Given $\{\mathbf{f}^i(\sigma)\}$ in \mathcal{H}_1 and constants \mathbf{c}^1 to \mathbf{c}^m problem (19) has a unique weak solution. If each $\mathbf{c}^i = 0$, the solution $\{\mathbf{r}^i\}$ belongs to \mathcal{H}_0^0 and satisfies*

$$\|\{\mathbf{r}^i(\sigma)\}\|_0 \leq C \|\{\mathbf{f}^i(\sigma)\}\|_1,$$

where C is a suitable constant. Since $\mathcal{H}_0^0 \subset \mathcal{H}_1$ we can define a linear transformation $S : \mathcal{H}_1 \rightarrow \mathcal{H}_1$, which maps the data $\{\mathbf{f}^i(\sigma)\}$ to the solution $\{\mathbf{r}^i(\sigma)\}$. S is self-adjoint, positive, injective and compact with dense image. Let \mathcal{A}^0 be defined as the inverse of S ; it is a self-adjoint operator densely defined in \mathcal{H}_1 with positive point spectrum

accumulating at ∞ . If $\{\mathbf{r}^i(\sigma)\}$ belongs to the domain of \mathcal{A}^0 , each $\mathbf{r}^i(\sigma)$ belongs to $W^2(0, \ell_i)$.

Now, motivated by (18), we define an unbounded operator \mathcal{B}^0 in \mathcal{H}_E^0 by

$$\text{dom}(\mathcal{B}^0) = \text{dom}(\mathcal{A}^0) \times \mathcal{H}_0 \quad \text{and} \quad \mathcal{B}^0 \begin{bmatrix} \{\mathbf{r}^{1,0}(\sigma)\} \\ \{\mathbf{r}^{i,1}(\sigma)\} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \mathcal{A}^0 & 0 \end{bmatrix} \begin{bmatrix} \{\mathbf{r}^{1,0}(\sigma)\} \\ \{\mathbf{r}^{i,1}(\sigma)\} \end{bmatrix}.$$

It turns out that \mathcal{B}^0 is skew-adjoint, so it generates a strongly continuous group $[U_t^0]_{t \in \mathbb{R}}$ of unitary operators on \mathcal{H}_E^0 . This allows us to obtain the solution $\{\mathbf{r}^i(\sigma, t)\}$ of (12), (13) (with $p = 0$), (14), (15), and (16) by taking the \mathcal{H}_0^0 - component of $U_t^0 \{\mathbf{r}^{i,0}(\sigma)\} \times \{\mathbf{r}^{i,1}(\sigma)\}$.

Next we need to treat the equations with inhomogeneous boundary conditions. When controls are imposed at the first p simple nodes while the remaining simple nodes are kept fixed, we expect the solutions to remain in the space \mathcal{H}_E^p . As a first step to solving the inhomogeneous boundary value problem, we extend the semigroup to \mathcal{H}_E from \mathcal{H}_E^0 . This involves identifying the orthogonal complement of \mathcal{H}_E^0 in \mathcal{H}_E with respect to an appropriately chosen inner product for the latter space. It turns out that this complement consists of stationary solutions of our linear system so that we can define the group $[U_t]_{t \in \mathbb{R}}$ of operators on \mathcal{H}_E by setting U_t equal to U_t^0 on \mathcal{H}_E^0 and equal to the identity on the orthogonal complement.

We note that on \mathcal{H}_0 (and, indeed, on the spaces \mathcal{H}_0^p with $p < m$), the inner product

$$(20) \quad \langle \{\mathbf{r}^i(\sigma)\}, \{\bar{\mathbf{r}}^i(\sigma)\} \rangle_0 + \sum_{i=1}^m \mathbf{r}^i(\ell_i) \cdot \bar{\mathbf{r}}^i(\ell_i),$$

where \cdot denotes the dot product of vectors, is equivalent to the standard inner product. Evidently, the orthogonal complement of the subspace \mathcal{H}_0^0 in \mathcal{H}_0 with respect to the latter inner product consists of all weak solutions of equations of the form (19) with $\{\mathbf{f}^i\} = \{0\}$; we denote that complement by \mathcal{S}_0 . We endow $\mathcal{H}_E = \mathcal{H}_0 \times \mathcal{H}_1$ with the corresponding “modified energy inner product.” This yields the desired orthogonal decomposition $\mathcal{H}_E = \mathcal{H}_E^0 \oplus \mathcal{S}_E$, where

$$\mathcal{S}_E = \{ \{\mathbf{r}^i(\sigma)\} \times \{0\} \mid \{\mathbf{r}^i(\sigma)\} \in \mathcal{S}_0 \}.$$

The orthogonal decomposition of data is performed as follows:

$$\{\mathbf{r}^{i,0}(\sigma)\} \times \{\mathbf{r}^{i,1}(\sigma)\} = \{\mathbf{r}^{i,0}(\sigma) - \bar{\mathbf{r}}^{i,0}(\sigma)\} \times \{\mathbf{r}^{i,1}(\sigma)\} \oplus \{\bar{\mathbf{r}}^{i,0}(\sigma)\} \times \{0\},$$

where $\mathcal{A}\{\bar{\mathbf{r}}^{i,0}(\sigma)\} = \{0\}$ and $\bar{\mathbf{r}}^{i,0}(\ell_i) = \mathbf{r}^{i,0}(\ell_i)$ for $i = 1$ to m . We easily check that the data in \mathcal{S}_E is stationary corresponding to nonhomogeneous, time independent boundary conditions at the simple nodes. It is therefore natural to define

$$U_t \{\mathbf{r}^{i,0}(\sigma)\} \times \{\mathbf{r}^{i,1}(\sigma)\} = U_t^0 \{\mathbf{r}^{i,0}(\sigma) - \bar{\mathbf{r}}^{i,0}(\sigma)\} \times \{\mathbf{r}^{i,1}(\sigma)\} \oplus \{\bar{\mathbf{r}}^i(\sigma)\} \times \{0\};$$

we thus obtain a group of unitary operators on \mathcal{H}_E . For initial data $\{\mathbf{r}^{i,0}(\sigma)\} \times \{\mathbf{r}^{i,1}(\sigma)\}$ in \mathcal{H}_E , we are allowed to solve the linear system with boundary conditions

$$\mathbf{r}^i(\ell_i, t) = \mathbf{r}^{i,0}(\ell_i) \text{ for } i = 1 \text{ to } m$$

by setting $\{\mathbf{r}^i(\sigma, t)\} \times \{\mathbf{r}^i(\sigma, t)\} = U_t \{\mathbf{r}^{i,0}(\sigma)\} \times \{\mathbf{r}^{i,1}(\sigma)\}$. We can also define unitary groups $[U_t^p]_{t \in \mathbb{R}}$ on \mathcal{H}_E^p by restriction (if $p < m$), or by taking quotients (if $p = m$).

We note that the generator \mathcal{B} of $[U_t]_{t \in R}$ is the direct sum $0 \oplus \mathcal{B}^0$ with domain $\mathcal{S}_E \oplus \text{dom}(\mathcal{B}^0)$. The generator \mathcal{B}^p of $[U_t^p]_{t \in R}$ is obtained from \mathcal{B} by restriction (if $p < m$) or by taking quotients (if $p = m$). If the initial data lies in the domain of \mathcal{B} for the solution obtained above we have that, for each i , $\mathbf{r}^i(\cdot, t)$ belongs to the spaces $C([0, T]; W^2(0, \ell_i))$, $C^1([0, T]; W^1(0, \ell_i))$ and $C^2([0, T]; L_2(0, \ell_i))$.

Now we consider the general initial boundary value problem (12)–(16), assuming that the initial data belongs to \mathcal{H}_E and that the following compatibility condition is satisfied:

$$(21) \quad \mathbf{r}^{i,0}(\ell_i) = \begin{cases} \mathbf{u}^i(0) & \text{for } i = 1 \text{ to } p, \\ 0 & \text{for } i = p + 1 \text{ to } m. \end{cases}$$

First, let $p = m$. We then obtain the solution as the sum of $U_t\{\mathbf{r}^{i,0}\} \times \{\mathbf{r}^{i,1}\}$ and the solution to the problem with homogeneous initial conditions and boundary conditions

$$\mathbf{r}^i(\ell_i, t) = \mathbf{u}^i(t) - \mathbf{u}^i(0) \text{ for } i = 1 \text{ to } m.$$

The approach to these inhomogeneous boundary conditions is standard: we move the inhomogeneity into the equation itself and then use the fact that when $\{\mathbf{f}^{i,0}(\sigma, t)\} \times \{\mathbf{f}^{i,1}(\sigma, t)\}$ is a continuous map of $[0, T]$ into $\mathcal{H}_E^0 = \mathcal{H}_0^0 \times \mathcal{H}_1$ the solution of

$$(22) \quad \frac{d}{dt} \begin{bmatrix} \{\mathbf{r}^i(\sigma, t)\} \\ \{\mathbf{r}_t^i(\sigma, t)\} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \mathcal{A} & 0 \end{bmatrix} \begin{bmatrix} \{\mathbf{r}^i(\sigma, t)\} \\ \{\mathbf{r}_t^i(\sigma, t)\} \end{bmatrix} + \begin{bmatrix} \{\mathbf{f}^{i,0}(\sigma, t)\} \\ \{\mathbf{f}^{i,1}(\sigma, t)\} \end{bmatrix},$$

subject to homogeneous boundary and initial conditions together with the coupling conditions, is given by

$$(23) \quad \{\mathbf{r}^i(\sigma, t)\} \times \{\mathbf{r}_t^i(\sigma, t)\} = \int_0^t U_{t-s}^0 \{\mathbf{f}^{i,0}(\sigma, s)\} \times \{\mathbf{f}^{i,1}(\sigma, s)\} ds,$$

which is again a continuous map of $[0, T]$ into $\mathcal{H}_E^0 = \mathcal{H}_0^0 \times \mathcal{H}_1$.

To transfer the inhomogeneity from the boundary condition into the equation, we begin by considering controls $\mathbf{u}^i(t) \in C_0^\infty((0, T])$. These functions vanish for $t = 0$ but not necessarily for $t = T$. For i between 1 and m , we then introduce functions $\phi^i(\sigma, t) \in C_0^\infty((0, \ell_i] \times (0, T])$ with $\phi^i(\ell_i, t) = \mathbf{u}^i(t)$; for $i > m$ it is convenient to set $\phi^i \equiv 0$. We now seek a solution to the inhomogeneous boundary value problem (with vanishing initial condition) in the form $\{\mathbf{r}^i(\sigma, t)\} = \{\phi^i(\sigma, t)\} + \{\bar{\mathbf{r}}^i(\sigma, t)\}$. Then $\{\bar{\mathbf{r}}^i(\sigma, t)\}$ will have to satisfy homogeneous boundary and initial conditions, the coupling conditions, as well as the equations

$$(24) \quad \rho_i \bar{\mathbf{r}}_{tt}^i = H^i \bar{\mathbf{r}}_{\sigma\sigma}^i + [H^i \phi_{\sigma\sigma}^i - \rho_i \phi_{t,t}^i] \text{ for } i = 1 \text{ to } n.$$

These equations can be written in the form (22) with $\mathbf{f}^{i,0}(\sigma, t)$ set equal to 0 and $\mathbf{f}^{i,1}(\sigma, t)$ equal to $H^i \phi_{\sigma\sigma}^i - \rho_i \phi_{t,t}^i$; this is then solved for $\{\bar{\mathbf{r}}^i(\sigma, t)\}$ using (23).

Applying results on the regularity of “mild solutions” to be found in Pazy [8] (in particular, Corollary 2.5 on p. 107), we can deduce that $\{\bar{\mathbf{r}}^i(\sigma, t)\} \times \{\bar{\mathbf{r}}_t^i(\sigma, t)\}$ is continuously differentiable with respect to t when regarded as a map from $[0, T]$ into \mathcal{H}_E^0 , and that this function and its t -derivatives take their values in $\text{dom}(\mathcal{B}^0)$. Consequently, each $\bar{\mathbf{r}}^i(\cdot, t)$ belongs to the spaces $C([0, T]; W^2(0, \ell_i))$, $C^1([0, T]; W^1(0, \ell_i))$ and $C^2([0, T]; L_2(0, \ell_i))$. That regularity is then inherited by $\{\mathbf{r}^i(\sigma, t)\}$.

System (12)–(16) can now also be solved for initial data in \mathcal{H}_E^p (with $p < m$) subject to the compatibility condition (21); the resulting trajectory remains in \mathcal{H}_E^p . Moreover, taking quotients, we can also solve the system in the space \mathcal{H}_E^m .

We have proved the following theorem.

THEOREM 3.2. *Given initial data $\{\mathbf{r}^{i,0}(\sigma)\} \times \{\mathbf{r}^{i,1}(\sigma)\}$ belonging to the space \mathcal{H}_E^p and boundary data $\mathbf{u}^1(t), \mathbf{u}^2(t), \dots, \mathbf{u}^p(t)$ belonging to $C^\infty([0, T])$ and satisfying the compatibility condition (21), the initial boundary value problem (12)–(16) has a solution such that $\{\mathbf{r}^i(\cdot, t)\} \times \{\mathbf{r}_t^i(\cdot, t)\}$ is a continuous trajectory in \mathcal{H}_E^p . Letting $[C_0^\infty((0, T))]^p$ denote the p -fold product of $C_0^\infty((0, T))$ we can define, for each $t \leq T$, a transformation $C_t^p : [C_0^\infty((0, T))]^p \rightarrow \mathcal{H}_E^p$ by assigning to the “control functions” $(\mathbf{u}^1, \mathbf{u}^2, \dots, \mathbf{u}^p)$ the solution at time t , namely $\{\mathbf{r}^i(\sigma, t)\} \times \{\mathbf{r}_t^i(\sigma, t)\}$, of the boundary value problem with zero initial values; in the case where $p = m$, this must be interpreted in terms of equivalence classes. The solution to the general initial boundary value problem is then given by*

$$(25) \quad \begin{aligned} \{\mathbf{r}^i(\cdot, t)\} \times \{\mathbf{r}_t^i(\cdot, t)\} &= U_t^p \{\mathbf{r}^{i,0}\} \times \{\mathbf{r}^{i,1}\} \\ &+ C_t^p(\mathbf{u}^1(\cdot) - \mathbf{u}^1(0), \mathbf{u}^2(\cdot) - \mathbf{u}^2(0), \dots, \mathbf{u}^p(\cdot) - \mathbf{u}^p(0)). \end{aligned}$$

If the initial data belongs to $\text{dom}(\mathcal{B})$, we have that each $\mathbf{r}^i(\cdot, t)$ belongs to the spaces $C([0, T]; W^2(0, \ell_i))$, $C^1([0, T]; W^1(0, \ell_i))$ and $C^2([0, T]; L_2(0, \ell_i))$.

Because of estimates in the next section, it will, in fact, be possible to extend the domain of the operators C_T^p from $[C_0^\infty((0, T))]^p$ to

$$\mathcal{U}^p = \{(\mathbf{u}^1, \mathbf{u}^2, \dots, \mathbf{u}^p) \in [W^1(0, T)]^p \mid \mathbf{u}^i(0) = 0 \text{ for } i = 1 \text{ to } p\}.$$

This latter space is the completion of $[C_0^\infty((0, T))]^p$ with respect to the norm associated with the inner product

$$(26) \quad \langle (\mathbf{v}^1, \mathbf{v}^2, \dots, \mathbf{v}^p), (\mathbf{u}^1, \mathbf{u}^2, \dots, \mathbf{u}^p) \rangle_{\mathcal{U}^p} = \frac{1}{2} \sum_{i=1}^p \int_0^T \mathbf{v}_t^i(t) \cdot \mathbf{u}_t^i(t) dt.$$

The operator C_T^p is studied thoroughly in the next section.

4. A priori estimates and the exact controllability of the linearized system. We are interested in the question of exact controllability for the control system (12)–(16). This involves establishing conditions on the network and on the time T , which ensure that the operator C_T^p becomes surjective, in which case any initial state in \mathcal{H}_E^p can be steered by suitable choice of the controls to any desired target state in the same space. First, we must specify the domain of C_T^p . Initially, we can regard C_T^p as a map from \mathcal{U}^p to \mathcal{H}_E^p having domain $[C_0^\infty((0, T))]^p$. Since the domain of C_T^p is dense, we can introduce the Hermitian adjoint $C_T^{p*} : \mathcal{H}_E^p \rightarrow \mathcal{U}^p$, which is identified in the next result involving “dual data” \mathbf{l}^i corresponding to a “dual problem.”

THEOREM 4.1. *Given $\{\mathbf{l}^{i,0}(\sigma)\} \times \{\mathbf{l}^{i,1}(\sigma)\}$ in \mathcal{H}_E^p , let $\{\mathbf{l}^i(\sigma, t)\}$ be a solution of the system (12) subject to the boundary conditions $\mathbf{l}^i(\ell_i, t) = \mathbf{l}^{i,0}(\ell_i)$ for $i = 1$ to m , the coupling conditions at the multiple nodes and the endpoint condition*

$$\{\mathbf{l}^i(\sigma, T)\} \times \{\mathbf{l}_T^i(\sigma, T)\} = \{\mathbf{l}^{i,0}(\sigma)\} \times \{\mathbf{l}^{i,1}(\sigma)\}.$$

Then, for each i from 1 to m , $\mathbf{l}_\sigma^i(\ell_i, t)$ is a well-defined function of t and there exists a constant K such that

$$(27) \quad \sum_{i=1}^m \int_0^T |H^i \mathbf{l}_\sigma^i(\ell_i, t)|^2 dt \leq K \| \{\mathbf{l}^{i,0}(\sigma)\} \times \{\mathbf{l}^{i,1}(\sigma)\} \|_E^2.$$

Moreover, $C_T^{p*} \{\mathbf{l}^{i,0}(\sigma)\} \times \{\mathbf{l}^{i,1}(\sigma)\}$ can be identified with $\{\mathbf{l}_\sigma^i(\ell_i, t) \frac{\partial}{\partial t}\}$, which acts on the controls as follows:

$$(28) \quad \begin{aligned} & \langle C_T^{p*} \{\mathbf{l}^{i,0}(\sigma)\} \times \{\mathbf{l}^{i,1}(\sigma)\}, (\mathbf{u}^1, \mathbf{u}^2, \dots, \mathbf{u}^p) \rangle_{\mathcal{U}^p} \\ &= \frac{1}{2} \sum_{i=1}^p \int_0^T H^i \mathbf{l}_\sigma^i(\ell_i, t) \cdot \mathbf{u}_t^i(t) dt. \end{aligned}$$

The proof of this theorem will be given later. First, we deduce some corollaries. We note that the right-hand side of (28) can be rewritten as

$$\begin{aligned} \frac{1}{2} \sum_{i=1}^p \int_0^T \left[\int_0^t H^i \mathbf{l}_\sigma^i(\ell^i, s) ds \right]_t \cdot \mathbf{u}_t^i(t) dt &= \left(\int_0^t H^1 \mathbf{l}_\sigma^1(\ell^1, s) ds, \int_0^t H^2 \mathbf{l}_\sigma^2(\ell^2, s) ds, \dots, \right. \\ &\quad \left. \int_0^t H^p \mathbf{l}_\sigma^p(\ell^p, s) ds \right), (\mathbf{u}^1, \mathbf{u}^2, \dots, \mathbf{u}^p) \rangle_{\mathcal{U}^p}. \end{aligned}$$

From the form of the inner product (26), we therefore get

$$(29) \quad \|C_T^{p*} \{\mathbf{l}^{i,0}(\sigma)\} \times \{\mathbf{l}^{i,1}(\sigma)\}\|_{\mathcal{U}^p}^2 = \sum_{i=1}^p \int_0^T |H^i \mathbf{l}_\sigma^i(\ell_i, t)|^2 dt.$$

This, together with (27), yields the following corollary.

COROLLARY 4.2. *The operator C_T^{p*} is bounded and, therefore, C_T^p can be extended to a bounded operator defined on the entire space \mathcal{U}^p .*

Now, by a general result on closed linear transformations, the surjectivity of $C_T^p : \mathcal{U}^p \rightarrow \mathcal{H}_E^p$ is equivalent to the estimate

$$\|C_T^{p*} \{\mathbf{l}^{i,0}(\sigma)\} \times \{\mathbf{l}^{i,1}(\sigma)\}\|_{\mathcal{U}^p} > k \|\{\mathbf{l}^{i,0}(\sigma)\} \times \{\mathbf{l}^{i,1}(\sigma)\}\|_{\mathcal{H}}$$

for some suitable positive constant k . In view of (29), we therefore obtain Corollary 4.3.

COROLLARY 4.3. *System (12)–(16) is exactly controllable in time T (i.e., $\text{im}(C_T^p) = \mathcal{H}_E^p$) if and only if the following a priori estimate is valid for some constant k :*

$$(30) \quad \sum_{i=1}^p \int_0^T |H^i \mathbf{l}_\sigma^i(\ell_i, t)|^2 dt \geq k \|\{\mathbf{l}^{i,0}(\sigma)\} \times \{\mathbf{l}^{i,1}(\sigma)\}\|_E^2.$$

Before proving Theorem 4.1, we derive a multiplier identity from which the estimates (27) and (30) can be deduced by the appropriate choice of multipliers. This approach to obtaining a priori estimate goes back to at least Rellich, and has recently proved very powerful in the control of partial differential equations; many applications can be found in Lions [6]. The following lemma applies to a generic string element, where the superscript has been dropped.

LEMMA 4.4. *Let $\mathbf{l}(\sigma, t)$ belong to $C([0, T]; W^2(0, \ell))$, $C^1([0, T]; W^1(0, \ell))$, and $C^2([0, T]; L_2(0, \ell))$ and be a solution of*

$$\rho \mathbf{l}_{tt}(\sigma, t) = H \mathbf{l}_{\sigma\sigma}(\sigma, t), \text{ for } (\sigma, t) \in [0, \ell] \times [0, T].$$

Let $M(\sigma)$ be a smooth, symmetric matrix-valued function that commutes with H for each σ . Then

$$(31) \quad \begin{aligned} \frac{1}{2} \int_0^T \left[H \mathbf{l}_\sigma \cdot M \mathbf{l}_\sigma \right]_{\sigma=0}^{\sigma=\ell} dt &= \frac{1}{2} \int_0^T \int_0^\ell [\rho \mathbf{l}_t \cdot M_\sigma \mathbf{l}_t + H \mathbf{l}_\sigma \cdot M_\sigma \mathbf{l}_\sigma] d\sigma dt \\ &+ \left[\int_0^\ell \rho \mathbf{l}_t \cdot M \mathbf{l}_\sigma d\sigma \right]_{t=0}^{t=T} - \frac{1}{2} \int_0^T \left[\rho \mathbf{l}_t \cdot M \mathbf{l}_t \right]_{\sigma=0}^{\sigma=\ell} dt. \end{aligned}$$

Proof. The proof is obtained by rewriting the following obvious identity:

$$\int_0^T \int_0^\ell [\rho \mathbf{l}_{tt} - H \mathbf{l}_{\sigma\sigma}] \cdot M \mathbf{l}_\sigma d\sigma dt = 0.$$

We have that

$$\begin{aligned} \int_0^T \int_0^\ell \rho \mathbf{l}_{tt} \cdot M \mathbf{l}_\sigma d\sigma dt &= \left[\int_0^\ell \rho \mathbf{l}_t \cdot M \mathbf{l}_\sigma d\sigma \right]_{t=0}^{t=T} - \int_0^\ell \int_0^T \rho \mathbf{l}_t \cdot M \mathbf{l}_{\sigma t} dt d\sigma \\ &= \left[\int_0^\ell \rho \mathbf{l}_t \cdot M \mathbf{l}_\sigma d\sigma \right]_{t=0}^{t=T} - \frac{1}{2} \int_0^T \left[\rho \mathbf{l}_t \cdot M \mathbf{l}_t \right]_{\sigma=0}^{\sigma=\ell} dt \\ &+ \frac{1}{2} \int_0^T \int_0^\ell \rho \mathbf{l}_t \cdot M_\sigma \mathbf{l}_t d\sigma dt; \end{aligned}$$

and

$$\begin{aligned} - \int_0^T \int_0^\ell H \mathbf{l}_{\sigma\sigma} \cdot M \mathbf{l}_\sigma d\sigma dt &= - \int_0^T \left[H \mathbf{l}_\sigma \cdot M \mathbf{l}_\sigma \right]_{\sigma=0}^{\sigma=\ell} dt \\ &+ \int_0^T \int_0^\ell [H \mathbf{l}_\sigma \cdot M \mathbf{l}_{\sigma\sigma} + H \mathbf{l}_\sigma \cdot M_\sigma \mathbf{l}_\sigma] d\sigma dt \\ &= - \frac{1}{2} \int_0^T \left[H \mathbf{l}_\sigma \cdot M \mathbf{l}_\sigma \right]_{\sigma=0}^{\sigma=\ell} dt + \int_0^T \int_0^\ell H \mathbf{l}_\sigma \cdot M_\sigma \mathbf{l}_\sigma d\sigma dt; \end{aligned}$$

here we have used $HM = MH$ and the symmetry of H and M to write

$$H \mathbf{l}_\sigma \cdot M \mathbf{l}_{\sigma\sigma} = \frac{1}{2} [H \mathbf{l}_\sigma \cdot M \mathbf{l}_\sigma]_\sigma - \frac{1}{2} H \mathbf{l}_\sigma \cdot M_\sigma \mathbf{l}_\sigma.$$

By adding and rearranging the last identities, we get (31).

Proof of Theorem 4.1. We prove estimate (27) for terminal data $\{\mathbf{l}^{i,0}\} \times \{\mathbf{l}^{i,1}\}$ belonging to the domain of \mathcal{B} (and to \mathcal{H}_E^p in the case where $p < m$) so that the regularity needed for Lemma 4.4 is assured; the general result follows by continuity. We then choose the multiplier $M(\sigma) = [-1 + (2/\ell)\sigma]I$ (with I the identity matrix). This satisfies $M(0) = -I$, $M(\ell) = I$ and $M_\sigma = 2\ell^{-1}I$; thus, discarding redundant terms of known sign, we easily get from (31)

$$\frac{1}{2} \int_0^T H \mathbf{l}_\sigma(\ell, t) \cdot \mathbf{l}_\sigma(\ell, t) \leq \frac{1}{2} \int_0^T \int_0^\ell [\rho \mathbf{l}_t \cdot \mathbf{l}_t + H \mathbf{l}_\sigma \cdot \mathbf{l}_\sigma] d\sigma dt + \left[\int_0^\ell \rho \mathbf{l}_t \cdot M \mathbf{l}_\sigma d\sigma \right]_{t=0}^{t=T}.$$

It easily follows that

$$\begin{aligned} \int_0^T |H\mathbf{l}_\sigma(\ell, t)|^2 dt &\leq A \int_0^T \int_0^\ell [\rho \mathbf{l}_t \cdot \mathbf{l}_t + H\mathbf{l}_\sigma \cdot \mathbf{l}_\sigma] d\sigma dt \\ &+ B \left[\int_0^\ell [\rho \mathbf{l}_t \cdot \mathbf{l}_t + H\mathbf{l}_\sigma \cdot \mathbf{l}_\sigma](\sigma, T) d\sigma + \int_0^\ell [\rho \mathbf{l}_t \cdot \mathbf{l}_t + H\mathbf{l}_\sigma \cdot \mathbf{l}_\sigma](\sigma, 0) d\sigma \right], \end{aligned}$$

where A and B are suitable constants. We can then write down this inequality for each string segment and sum these inequalities to obtain

$$\begin{aligned} \sum_{i=1}^n \int_0^T |H^i \mathbf{l}_\sigma^i(\ell_i, t)|^2 dt &\leq A \int_0^T \|\{\mathbf{l}^i(\sigma, t)\} \times \{\mathbf{l}_t^i(\sigma, t)\}\|_E^2 dt \\ &+ B [\|\{\mathbf{l}^{i,0}(\sigma)\} \times \{\mathbf{l}^{i,1}(\sigma)\}\|_E^2 + \|\{\mathbf{l}^i(\sigma, 0)\} \times \{\mathbf{l}_t^i(\sigma, 0)\}\|_E^2], \end{aligned}$$

where A and B have been suitably redefined. By using conservation of energy, we have that $\|\{\mathbf{l}^i(\sigma, t)\} \times \{\mathbf{l}_t^i(\sigma, t)\}\|_E^2 = \|\{\mathbf{l}^{i,0}(\sigma)\} \times \{\mathbf{l}^{i,1}(\sigma)\}\|_E^2$ for all t so that (27) holds with $K = AT + B$. These constants can be estimated explicitly in terms of the various physical parameters. We note that, had we not discarded so many terms, we would have obtained the following generalisation of (27):

$$(32) \quad \sum_{i=10}^n \int_0^T [H^i \mathbf{l}_\sigma^i(\ell_i, t)|^2 + |H^i \mathbf{l}_\sigma^i(0, t)|^2 + |\mathbf{l}_t^i(\ell_i, t)|^2 + |\mathbf{l}_t^i(0, t)|^2] dt \leq K \|\{\mathbf{l}^{i,0}(\sigma)\} \times \{\mathbf{l}^{i,1}(\sigma)\}\|_E^2.$$

To prove (28) it is enough to establish
(33)

$$\langle \{\mathbf{l}^{i,0}(\sigma)\} \times \{\mathbf{l}^{i,1}(\sigma)\}, C_T^p(\mathbf{u}^1, \mathbf{u}^2, \dots, \mathbf{u}^p) \rangle_{\mathcal{U}^p} = \frac{1}{2} \sum_{i=1}^p \int_0^T H^i \mathbf{l}_\sigma^i(\ell_i, t) \cdot \frac{\partial}{\partial t} \mathbf{u}^i(t) dt.$$

Let $\{\mathbf{r}^i(\sigma, t)\}$ be the solution of (12) involved in the definition of C_T^p and $\{\mathbf{l}^i(\sigma, t)\}$ be as above. Then

$$\begin{aligned} 2 \frac{d}{dt} \langle \{\mathbf{r}^i(\sigma, t)\} \times \{\mathbf{r}_t^i(\sigma, t)\}, \{\mathbf{l}^i(\sigma, t)\} \times \{\mathbf{l}_t^i(\sigma, t)\} \rangle_E \\ = \sum_{i=1}^n \int_0^{\ell_i} [\rho_i \mathbf{r}_{tt}^i \cdot \mathbf{l}_t^i + \rho_i \mathbf{r}_t^i \cdot \mathbf{l}_{tt}^i + H^i \mathbf{r}_{\sigma t}^i \cdot \mathbf{l}_\sigma^i + H^i \mathbf{r}_\sigma^i \cdot \mathbf{l}_{\sigma t}^i] d\sigma \\ = \sum_{i=1}^n \left[H^i \mathbf{r}_t^i \cdot \mathbf{l}_\sigma^i + H^i \mathbf{r}_\sigma^i \cdot \mathbf{l}_t^i \right]_{\sigma=0}^{\sigma=\ell_i}. \end{aligned}$$

By integrating from 0 to T and using the initial boundary and coupling conditions (which involves extensive regrouping of the boundary terms corresponding to multiple nodes), we obtain (33), and this completes the proof of Theorem 4.1.

Finally, we return to the question of exact controllability. By Corollary 4.3 we need to establish the estimate (30); for this the configuration of the network plays an important role. Again, we can assume that the terminal data belongs to the domain of \mathcal{B}^p and obtain the general estimate by continuity. We now use multipliers of the

form $M^i(\sigma) = (\alpha_i\sigma + \beta_i)I$ (with β_i positive) so that $M_\sigma^i = \alpha_i I$ and $M^i(0) = \beta_i I$. Adding the identities (31) over all string elements we get on rearrangement

$$\begin{aligned}
 & \frac{1}{2} \sum_{i=1}^p \int_0^T [H^i \mathbf{l}_\sigma^i \cdot M^i \mathbf{l}_\sigma^i](\ell_i, t) dt \\
 &= \sum_{i=1}^n \beta_i \frac{1}{2} \int_0^T \int_0^{\ell_i} [\rho_i \mathbf{l}_t^i \cdot \mathbf{l}_t^i + H^i \mathbf{l}_\sigma^i \cdot \mathbf{l}_\sigma^i] d\sigma dt \\
 &+ \sum_{i=1}^n \left[\int_0^\ell \rho_i \mathbf{l}_t^i \cdot M^i \mathbf{l}_\sigma^i d\sigma \right]_{t=0}^{t=T} - \frac{1}{2} \sum_{i=p+1}^m \int_0^T [H^i \mathbf{l}_\sigma^i \cdot M^i \mathbf{l}_\sigma^i](\ell_i, t) dt \\
 &- \frac{1}{2} \sum_{\mathbf{N} \in \mathcal{N}_m} \sum_{i \in \mathcal{E}(\mathbf{N})} \epsilon_i(\mathbf{N}) \int_0^T [H^i \mathbf{l}_\sigma^i \cdot M^i \mathbf{l}_\sigma^i](\mathbf{N}, t) dt \\
 (34) \quad &- \frac{1}{2} \sum_{\mathbf{N} \in \mathcal{N}_m} \sum_{i \in \mathcal{E}(\mathbf{N})} \epsilon_i(\mathbf{N}) \int_0^T [\rho_i \mathbf{l}_t^i \cdot M^i \mathbf{l}_t^i](\mathbf{N}, t) dt,
 \end{aligned}$$

where \mathcal{N}_m is the set of multiple nodes, and we have used the fact that $\mathbf{l}_t^i(\ell_i, t)$ vanishes for $i = 1$ to m . By conservation of energy and Schwarz's inequality,

$$(35) \quad \sum_{i=1}^n \frac{1}{2} \int_0^T \int_0^{\ell_i} [\rho_i \mathbf{l}_t^i \cdot \mathbf{l}_t^i + H^i \mathbf{l}_\sigma^i \cdot \mathbf{l}_\sigma^i] d\sigma dt = T \| \{ \mathbf{l}^{i,0}(\sigma) \} \times \{ \mathbf{l}^{i,1}(\sigma) \} \|_E^2$$

and

$$(36) \quad \sum_{i=1}^n \left[\int_0^\ell \rho_i \mathbf{l}_t^i \cdot M^i \mathbf{l}_\sigma^i d\sigma \right]_{t=0}^{t=T} \geq \delta \| \{ \mathbf{l}^{i,0}(\sigma) \} \times \{ \mathbf{l}^{i,1}(\sigma) \} \|_E^2,$$

where δ is a positive constant. If

$$(37) \quad M^i(\ell_i) > 0 \quad \text{for } i = 1 \text{ to } p,$$

then we readily obtain from (34), using (35) and (36), that

$$\begin{aligned}
 & \gamma \sum_{i=1}^p \int_0^T |H^i \mathbf{l}_\sigma^i(\ell_i, t)|^2 dt \geq \alpha(T - T_0) \| \{ \mathbf{l}^{i,0}(\sigma) \} \times \{ \mathbf{l}^{i,1}(\sigma) \} \|_E^2 \\
 &- \frac{1}{2} \sum_{i=p+1}^m \int_0^T [H^i \mathbf{l}_\sigma^i \cdot M^i \mathbf{l}_\sigma^i](\ell_i, t) dt - \frac{1}{2} \sum_{\mathbf{N} \in \mathcal{N}_m} \sum_{i \in \mathcal{E}(\mathbf{N})} \epsilon_i(\mathbf{N}) \\
 &\cdot \int_0^T [H^i \mathbf{l}_\sigma^i \cdot M^i \mathbf{l}_\sigma^i](\mathbf{N}, t) dt - \frac{1}{2} \sum_{\mathbf{N} \in \mathcal{N}_m} \sum_{i \in \mathcal{E}(\mathbf{N})} \epsilon_i(\mathbf{N}) \int_0^T [\rho_i \mathbf{l}_t^i \cdot M^i \mathbf{l}_t^i](\mathbf{N}, t) dt,
 \end{aligned}$$

with α the minimum of the α_i 's, $T_0 = \delta/\alpha$, and γ a positive number such that $\gamma H^i \leq M^i(\ell_i)$ for $i = 1$ to p . We finally obtain (30) with $k = \alpha/\gamma(T - T_0)$ if the configuration is such that the multipliers can be chosen to satisfy (37) along with

$$(38) \quad M^i(\ell_i) \leq 0 \quad \text{for } i = p+1 \text{ to } m;$$

$$(39) \quad \sum_{\mathbf{N} \in \mathcal{N}_m} \sum_{i \in \mathcal{E}(\mathbf{N})} \epsilon_i(\mathbf{N}) [H^i \mathbf{l}_\sigma^i \cdot M^i \mathbf{l}_\sigma^i](\mathbf{N}, t) \leq 0;$$

and

$$(40) \quad \sum_{\mathbf{N} \in \mathcal{N}_m} \sum_{i \in \mathcal{E}(\mathbf{N})} \epsilon_i(\mathbf{N}) [\rho_i \mathbf{l}_t^i \cdot M^i \mathbf{l}_t^i](\mathbf{N}, t) \leq 0.$$

By the first of the coupling conditions, $\mathbf{l}_t^i(\mathbf{N}, t)$ is the same for each $i \in \mathcal{E}(\mathbf{N})$ and so (40) can be replaced by the condition

$$(41) \quad \sum_{i \in \mathcal{E}(\mathbf{N})} \epsilon_i(\mathbf{N}) \rho_i M^i(\mathbf{N}) \quad \text{is negative definite for each } \mathbf{N} \in \mathcal{N}_m.$$

The second coupling condition can also be used to rewrite (39) as

$$(42) \quad \sum_{i \in \mathcal{E}(\mathbf{N})} \mathbf{w}^i = 0 \quad \text{implies} \quad \sum_{i \in \mathcal{E}(\mathbf{N})} \epsilon_i(\mathbf{N}) M^i(\mathbf{N}) H^{i-1} \mathbf{w}^i \cdot \mathbf{w}^i \leq 0.$$

It is convenient to introduce $\overline{M}^i(\mathbf{N}) = \epsilon_i(\mathbf{N}) M^i(\mathbf{N})$, which at the simple nodes coincides with $M^i(\mathbf{N})$. Condition (41) is then satisfied if at every multiple node at least one $\overline{M}^i(\mathbf{N})$ is sufficiently negative to dominate the positive contributions at that node. It is an exercise in linear algebra to see that the subtler condition (42) will be satisfied if at each internal node at most one $\overline{M}^i(\mathbf{N})$ is positive and if each negative contribution is sufficiently small to dominate the positive contribution.

For a given network we try to define appropriate multipliers satisfying the above conditions. As a first step, we proceed “graphically” as follows. String segments will be denoted by

$$\bullet - \pm - \pm - \bullet,$$

where the signs indicate the signs of $\overline{M}^i(\mathbf{N})$ at the adjacent node. We note that because $M^i(\sigma)$ is monotone increasing the only possible sign allocations are

$$(43) \quad \bullet - + - + - \bullet, \quad \bullet - + - - - \bullet, \quad \text{and} \quad \bullet - - - + - \bullet;$$

the assignment $\bullet - - - - \bullet$ is impossible. Subject to this “constraint” we try to satisfy the following *sign rules*, which are necessary for our approach to yield exact controllability:

- (a) At the simple nodes we must assign positive values where controls are to be applied and negative values where the node is to be held fixed;
- (b) At each multiple node we must assign positive or negative values in such a way that at most one is positive.

If such a sign allocation is possible, we proceed to check whether the magnitudes of $\overline{M}^i(\mathbf{N})$ at both ends of each element can be adjusted to give the required dominance of the negative terms over the positive terms needed to ensure that (41) and (42) hold.

In the networks illustrated on the next page, the simple nodes at which controls are applied are marked by “ \mathbf{u}^i ”, while those that are kept fixed are marked by “0.”

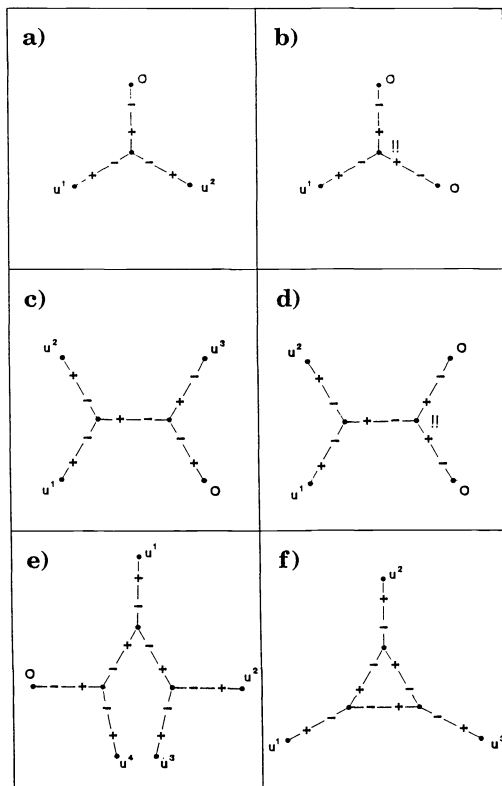


FIG. 2

Nodes where the “sign rule” is violated are indicated by “!!,” and, in the cases of violation, we cannot remedy the situation by a different sign allocation. With these remarks, the diagrams should be self explanatory. When we try to adjust the values of $\overline{M}^i(\mathbf{N})$ for the networks in which the sign rules can be satisfied, we see that this is easily done when there are no “closed circuits.” Consequently, we have established the following theorem.

THEOREM 4.5. *The networks (a), (c) and (e) illustrated below in Fig. 2 can all be exactly controlled by controls at the indicated simple nodes.*

Remarks

- In the networks (b) and (d) where our sign rules cannot be satisfied, we are unable to make any general assertions; in the example given below, exact controllability of the network depends on the rationality or irrationality of a certain parameter. It appears that in some situations there is no interaction between certain oscillations occurring in part of the network and the oscillations on the rest of the network excited by control functions at some set of simple nodes.
- In the case of the network (f) with its closed circuit, we can easily convince ourselves that we cannot ensure that condition (42) is satisfied at all multiple nodes. This problem persists in more complicated networks involving closed circuits.

Whether such circuits are exactly controllable remains open.

- (c) We can easily generate exactly controllable networks that are much more complicated than those that we have treated.

Example. We consider network (b) with boundary conditions

$$\mathbf{r}^1(\ell_1, t) = \mathbf{u}(t), \quad \mathbf{r}^2(\ell_2, t) = \mathbf{0}, \quad \mathbf{r}^3(\ell_3, t) = \mathbf{0}.$$

We suppose for simplicity that the second and third strings are identical, perpendicular to each other, and lie symmetrically with respect to the third string. Moreover, we set $\rho_i = h_i = 1$ for $i = 2, 3$ and note that in this case $s_2 = s_3 = s$, say. We introduce the parameter μ by $\mu^{-2} = 1 - s^{-1}$. If μ is rational we can find two nontrivial real functions f and g defined on the real line that are both 2 and 2μ periodic and satisfy $f(t) = -g(-t)$. It follows easily that

$$f(\mu + t) + g(\mu - t) = 0 \quad \text{and} \quad f(1 + t) + g(1 - t) = 0.$$

Now let \mathbf{v} be the direction vector of the second string and set

$$\begin{aligned} \mathbf{l}^1(\sigma, t) &= \mathbf{0}, \\ \mathbf{l}^2(\sigma, t) &= [f(\sigma + t) + g(\sigma - t)]\mathbf{v}, \\ \mathbf{l}^3(\sigma, t) &= -\mu[f(\mu\sigma + t) + g(\mu\sigma - t)]\mathbf{v}. \end{aligned}$$

This gives a solution $\{\mathbf{l}^i(\sigma, t)\}$ to system (12) satisfying homogeneous boundary conditions as well as the two coupling conditions (14) and (15). Moreover, in this situation $C_T^{p*} \{\mathbf{l}^i(\sigma, T)\} \times \{\mathbf{l}_t^i(\sigma, T)\} = 0$, so that C_T^{p*} has a nontrivial nullspace, which implies that C_T^p is not surjective, and hence the system is not exactly controllable.

With an additional argument, we can show that when μ is irrational, C_T^{p*} is injective for T sufficiently large, in which case C_T^p has dense image. At this point we do not know whether the image is necessarily closed. Questions raised by the above example could be the topic of a future publication.

REFERENCES

- [1] S. S. ANTMAN, *The equations for large vibrations of strings*, Math. Monthly, 87 (1980), pp. 359–370.
- [2] G. F. CARRIER, *On the non-linear vibration problem of the elastic string*, Quart. Appl. Math., 2, (1945), pp. 157–165.
- [3] 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.
- [4] J. LAGNESE, *Decay of solutions of wave equations with boundary dissipation*, J. Differential Equations, 50 (1983), pp. 163–182.
- [5] 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, 3, pp. 2287–2290.
- [6] J. L. LIONS, *Cóntrollabilité Exacte, Perturbations et Stabilisation de Systèmes Distribués*, Tomes 1 et 2, Masson, Paris, 1988.
- [7] J. NEČAS, *Les Méthodes Directes en Théorie des Equations Elliptiques*, Academia, Prague and Masson et Cie, Paris, (1967).

- [8] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, New York, 1983.
- [9] E. J. P. G. SCHMIDT, *Exact controllability of vibrating strings—a preliminary report*, Report No. 112, Schwerpunktprogramm der Deutschen Forschungsgemeinschaft: Anwendungsbezogene Optimierung und Steuerung, Technische Hochschule Darmstadt, 1989.

ON THE EXISTENCE OF OPTIMAL RELAXED CONTROLS OF STOCHASTIC PARTIAL DIFFERENTIAL EQUATIONS*

XUN YU ZHOU†

Abstract. This paper is concerned with control problem of systems governed by stochastic partial differential equations, the drift and diffusion terms of which are second- and first-order differential operators, respectively. The existence of an optimal relaxed control is studied in both cases where the systems are degenerate and nondegenerate. It is shown that the higher regularity conditions on the initial state, as required in the existing results, can be dispensed with if the Wiener process is one-dimensional. Some special cases of multidimensional Wiener process are also discussed, which in particular leads to an improvement of a recent result of Bensoussan and Nisio. The method is based on an analysis of the group generated by the first-order differential operator. As an application, an existence theorem of the optimal relaxed control is proved for partially observed diffusions with correlation between the controlled states and the observation noises.

Key words. stochastic partial differential equation (SPDE), optimal relaxed control, partially observed diffusion, group of operators

AMS(MOS) subject classifications. 60H15, 93E20, 49A60, 93E11

1. Introduction. In this paper we consider an optimal control problem of the following kind of stochastic partial differential equations (SPDE):

$$\begin{aligned}
 dq(t, x) = & [\partial_i(a^{ij}(x, u(t))\partial_j q(t, x)) + b^i(x, u(t))\partial_i q(t, x) \\
 & + c(x, u(t))q(t, x) + f(x, u(t))] dt \\
 & + [\sigma^i(x)\partial_i q(t, x) + h(x)q(t, x) + g(x)] dW(t), \\
 & x \in R^d, \quad t \in [0, T], \\
 q(0, x) = & q_0(x), \quad x \in R^d,
 \end{aligned}
 \tag{1.1}$$

where W is a one-dimensional Wiener process with $W(0) = 0$, $\{u(t): 0 \leq t \leq T\}$ is an admissible control (in the usual sense), and $\partial_i := \partial/\partial x_i$, $i = 1, 2, \dots, d$. Note that here and in the following we always use the conventional repeated indices for summation.

The optimal control problem is to minimize a given cost functional over the totality of admissible controls. SPDE (1.1) occurs in many areas of science, especially in physics (see [8]–[10] and [13] and the references therein). From the mathematical point of view, the most important example is perhaps the filtering problem. More precisely, the control problem of partially observed diffusions can be reduced to the control problem of SPDE (1.1) (Zakaï's equation), with the σ^i in (1.1) corresponding to the correlation between the controlled state and the observation noises ([12], [13], and [15]).

The existence of an optimal control in the usual sense seems to be a very difficult problem and remains open in general. Many authors turned to studying the relaxed control problem ([1], [4], [6] and [13]). For system (1.1), when $\sigma^i = 0$, Bensoussan and Nisio [1] have proved the existence of an optimal relaxed control, assuming (a^{ij}) is uniformly positive definite and $q_0 \in H^2(R^d)$ ($H^k(R^d) :=$ Sobolev space $W_2^k(R^d)$). When $\sigma^i \neq 0$, the situation is much more complicated. For this case, the recent delicate

* Received by the editors April 9, 1990; accepted for publication (in revised form) October 20, 1990. This research was partially supported by the Monbusho Scholarship of Japanese Government and the National Natural Science Foundation of China.

† Department of Mathematics, Faculty of Science, Kobe University, Rokko, Kobe 657, Japan.

work of Nagase and Nisio [13] shows that the existence theorem still holds when $(a^{ij} - 3/2\sigma^i\sigma^j)$ is nonnegative definite and $q_0 \in H^3(R^d)$. Both [1] and [13] have applied a similar method to that of Nagase [12], the key point of which is to employ a compact embedding lemma, see [12, Remark 3.1]. These results, however, require the higher regularity on the initial point q_0 .

The purpose of the present paper is to establish the existence of an optimal relaxed control of SPDE (1.1), with some natural regularity conditions on the initial state q_0 . Different from that of [1], [12], and [13], our method depends on an analysis of the group generated by the first-order differential operator, inspired by Da Prato, Iannelli, and Tubaro [3]. This frees us from the set-up of [1], [12], and [13], allowing us to eliminate the higher regularity restriction on q_0 . The main results of this paper, roughly speaking, are as follows: Either

- (i) $(a^{ij} - 1/2\sigma^i\sigma^j)$ is nonnegative definite and $q_0 \in H^1(R^d)$, or
- (ii) $(a^{ij} - 1/2\sigma^i\sigma^j)$ is uniformly positive definite and $q_0 \in H^0(R^d) (= L^2(R^d))$

will ensure the existence of an optimal relaxed control.

An apparent defect of our method, however, is that the Wiener process W is required to be one-dimensional. For multidimensional cases, the method applies only to some special cases. In particular, we will show that our method is effective to the setting of Bensoussan and Nisio [1]; therefore, their result may be considerably improved.

It is also worth noting that this work benefits so much by the delicate and deep results of Krylov and Rozovskii [8]–[10] concerning the SPDE theory, and in particular, the a priori estimate of differential operators (Lemma 3.1 below).

The paper is organized as follows: In § 2, we will give a precise definition of the relaxed system of (1.1) as well as some basic notation and facts. Sections 3 and 4 are the main parts of the paper, in which the existence theorems are proved for degenerate case (i.e., $(a^{ij} - 1/2\sigma^i\sigma^j)$ is nonnegative definite) and nondegenerate case (i.e., $(a^{ij} - 1/2\sigma^i\sigma^j)$ is uniformly positive definite), respectively. In § 5, we discuss the cases when the Wiener process is multidimensional, as well as the application of the main results to the partially observed diffusions.

2. Preliminaries. We define a family of second-order differential operators $\{A(u): u \in \Gamma \subset R^m\}$ and a first-order differential operator M by

$$(2.1) \quad A(u)\phi(x) := \partial_i(a^{ij}(x, u)\partial_j\phi(x)) + b^i(x, u)\partial_i\phi(x) + c(x, u)\phi(x),$$

$$(2.2) \quad M\phi(x) := \sigma^i(x)\partial_i\phi(x) + h(x)\phi(x), \quad \text{for } x \in R^d,$$

where a^{ij} , b^i , c , σ^i , and h are real-valued functions for $i, j = 1, 2, \dots, d$.

Let H^k be the Sobolev space $W_2^k(R^d)$ with the norm $\|\cdot\|_k$ ($k = 0, \pm 1, \pm 2, \dots$). $\langle \cdot, \cdot \rangle_k$ denotes the duality pairing between H^{k-1} and H^{k+1} under $(H^k)^* = H^k$, and $(\cdot, \cdot)_k$ is the inner product in H^k . For $r \geq 0$, define

$$L_r^2 := \{\phi: \phi \text{ is a real-valued Borel function on } R^d,$$

$$\text{and } (1 + |\cdot|^2)^{r/2}\phi(\cdot) \in H^0\},$$

with the norm $\|\phi\|_{0,r} := (\int_{R^d} |(1 + |x|^2)^{r/2}\phi(x)|^2 dx)^{1/2}$.

Let H_r^k be the subspace of L_r^2 consisting of functions whose generalized derivatives up to the order k belong to L_r^2 . It becomes a Hilbert space with the norm

$$\|\phi\|_{k,r} := \left(\sum_{|\alpha| \leq k} \frac{|\alpha|!}{\alpha^1! \cdots \alpha^d!} \|D^\alpha \phi\|_{0,r}^2 \right)^{1/2},$$

see [10, § 2].

For any second-order differential operator L , when we write $\langle L\phi, \psi \rangle_k$, then L is understood to be an operator from H^{k+1} to H^{k-1} by formally using Green's formula. For example, for the operators $A(u)$ defined by (2.1), we have

$$(2.3) \quad \begin{aligned} \langle A(u)\phi, \psi \rangle_k &:= -(a^{ij}(\cdot, u)\partial_j\phi, \partial_i\psi)_k + (b^i(\cdot, u)\partial_i\phi, \psi)_k \\ &\quad + (c(\cdot, u)\phi, \psi)_k \quad \text{for } \phi, \psi \in H^{k+1}. \end{aligned}$$

Now we recall the definition of relaxed controls, according to [1], [4], and [13]. By Λ we denote the set of all measures λ on $[0, T] \times \Gamma$ such that

$$(2.4) \quad \lambda([0, s] \times \Gamma) = s, \quad \text{for } s \leq T.$$

If Γ is a compact set, then Λ is compact when endowed with the weak convergence topology ([1] and [13]).

Set $\sigma_t(\Lambda) :=$ the σ -field generated by $\{\lambda: \lambda([0, s] \times A) \in \mathcal{B}(R^+), s \leq t, A \in \mathcal{B}(\Gamma)\}$ and $\sigma(\Lambda) := \sigma_T(\Lambda)$. Let $\mathcal{P} := \mathcal{P}(\Lambda)$ be the space of probabilities on $(\Lambda, \sigma(\Lambda))$, then Prohorov's theorem yields that \mathcal{P} is a compact metric space when endowed with the weak convergence topology.

By (2.4), λ is represented by $\lambda(dt, du) = \lambda'(t, du) dt$, where $\lambda'(t, \cdot)$ is a probability on Γ for almost all t and determined uniquely except t -null set. For any bounded and uniformly continuous function ρ on $R^d \times \Gamma$, set $\tilde{\rho}(t, x, \lambda) := \int_{\Gamma} \rho(x, u) \lambda'(t, du)$. Define a family of operators $\{\tilde{A}(t, \lambda): t \in [0, T], \lambda \in \Lambda\}$:

$$(2.5) \quad \begin{aligned} \tilde{A}(t, \lambda)\phi(x) &:= \partial_i(\tilde{a}^{ij}(t, x, \lambda)\partial_j\phi(x)) + \tilde{b}^i(t, x, \lambda)\partial_i\phi(x) \\ &\quad + \tilde{c}(t, x, \lambda)\phi(x), \quad \text{for } x \in R^d. \end{aligned}$$

Now we introduce the relaxed system.

DEFINITION 2.1. $\mathcal{R} = (\Omega, \mathcal{F}, P, \mathcal{F}_t, W, \mu)$ is called a relaxed system if

$$(2.6) \quad (\Omega, \mathcal{F}, P, \mathcal{F}_t) \text{ is a standard probability space with filtration } \{\mathcal{F}_t: 0 \leq t \leq T\};$$

$$(2.7) \quad W \text{ is an } \mathcal{F}_t\text{-adapted one-dimensional Wiener process with } W(0) = 0;$$

$$(2.8) \quad \mu \text{ is an } \mathcal{F}_t\text{-adapted } \Lambda\text{-valued random variable } (\Lambda\text{-r.v.}), \text{ i.e., } \mu(B_1 \times B_2) \text{ is } \mathcal{F}_t\text{-measurable whenever } B_1 \in \mathcal{B}([0, t]) \text{ and } B_2 \in \mathcal{B}(\Gamma).$$

For simplicity, we put $\mathcal{R} = (W, \mu)$ if no confusion arises, and sometimes we simply call μ a relaxed control.

\mathcal{R} denotes the totality of relaxed controls. For $\mathcal{R} = (W, \mu)$, $\pi(\mathcal{R})$ denotes the image measure of (W, μ) on $C(0, T; R^1) \times \Lambda$. Again, by endowing the space $\Pi := \{\pi(\mathcal{R}): \mathcal{R} \in \mathcal{R}\}$ with the weak convergence topology, we have the following proposition [1], [13].

PROPOSITION 2.1. Π is a compact metric space.

DEFINITION 2.2. We say \mathcal{R}_n converges to \mathcal{R} , writing $\mathcal{R}_n \rightarrow \mathcal{R}$, if $\pi(\mathcal{R}_n) \rightarrow \pi(\mathcal{R})$ weakly.

Given $\mathcal{R} = (\Omega, \mathcal{F}, P, \mathcal{F}_t, W, \mu)$, consider the following SPDE:

$$(2.9) \quad \begin{aligned} dq(t) &= (\tilde{A}(t, \mu)q(t) + \tilde{f}(t, \mu)) dt + (Mq(t) + g) dW(t), \\ q(0) &= q_0. \end{aligned}$$

An H^1 -valued \mathcal{F}_t -adapted process $q = q^{\mathcal{R}}(\cdot, q_0)$ is called a solution of (2.9) or a response for the relaxed control \mathcal{R} if

$$(2.10) \quad E \int_0^T \|q(t)\|_1^2 dt < +\infty,$$

and for any $\eta \in C_0^\infty(R^d)$ (smooth function on R^d with compact support) and almost

all $(t, \omega) \in [0, T] \times \Omega$,

$$(2.11) \quad \begin{aligned} (q(t), \eta)_0 &= (q_0, \eta)_0 + \int_0^t \langle \tilde{A}(s, \mu)q(s), \eta \rangle_0 ds + \int_0^t (\tilde{f}(s, \mu), \eta)_0 ds \\ &\quad + \int_0^t (Mq(s) + g, \eta)_0 dW(s). \end{aligned}$$

For each initial q_0 , we are given a cost functional

$$(2.12) \quad J(q_0, \mathcal{R}) := E\{F(q^{\mathcal{R}}(\cdot, q_0)) + G(q^{\mathcal{R}}(T, q_0))\}, \quad \mathcal{R} \in \mathbf{R}.$$

The optimal relaxed control problem is to minimize $J(q_0, \cdot)$ over \mathbf{R} , for each q_0 .

Remark 2.1. Since we will mainly consider the relaxed control in the following, we will simply write $A(t, \mu) = \tilde{A}(t, \mu)$, etc., if no confusion arises.

3. Degenerate case. Let us fix a positive constant K . We introduce the following conditions on the functions appearing in (1.1):

- (A1) $a^{ij}, b^i, c: R^d \times \Gamma \rightarrow R^1, \sigma^i, h: R^d \rightarrow R^1$ are continuous functions; these functions and their derivatives in x up to second order do not exceed K in absolute value;
- (A2) $a^{ij} = a^{ji}, i, j = 1, 2, \dots, d$, and $(a^{ij} - \frac{1}{2}\sigma^i\sigma^j)_{ij}$ is a nonnegative definite matrix;
- (A3) $f(\cdot, u) \in H^1, g \in H^2$; the absolute values of f, g together with the H^1 -norm of $f(\cdot, u)$ and the H^2 -norm of g do not exceed K ;
- (A3)_r For some $r > 0, f(\cdot, u) \in L_r^2, g \in H_r^1$, the absolute values of f, g together with the L_r^2 -norm of $f(\cdot, u)$ and the H_r^1 -norm of g do not exceed K .

The following lemma of a prior estimate is a special case of [9, Lemma 2.1].

LEMMA 3.1. *Let \hat{A} and \hat{M} be any second-order and first-order differential operators that have the forms of (2.1) and (2.2), respectively, and whose coefficients satisfy (A1) and (A2). Then there exists a constant N depending only on K in (A1) such that*

$$(3.1) \quad 2\{\langle \hat{A}\phi, \phi \rangle_k + \langle \hat{M}\phi, \phi \rangle_k\} + \|\hat{M}\phi + \hat{g}\|_k^2 \leq N(\|\phi\|_k^2 + \|\hat{f}\|_k^2 + \|\hat{g}\|_{k+1}^2),$$

for any $\phi \in H^{k+1}, \hat{f} \in H^k, \hat{g} \in H^{k+1}; k = 0, 1$.

COROLLARY 3.1. *Let \hat{M} be a first-order differential operator that has the form of (2.2) and whose coefficients satisfy (A1). Then there exists a constant N depending only on K such that*

$$(3.2) \quad |\langle \hat{M}\phi + \hat{g}, \phi \rangle_k| \leq N(\|\phi\|_k^2 + \|\hat{g}\|_k^2), \quad \text{for any } \phi \in H^{k+1}, \hat{g} \in H^k; k = 0, 1.$$

The following result concerning the solution of SPDE is known from Krylov and Rozovskii [9] and [10].

PROPOSITION 3.1. *Assume (A1)–(A3) and $q_0 \in H^1$; then for any $\mathcal{R} \in \mathbf{R}$, (2.9) has a unique solution $q^{\mathcal{R}}(\cdot, q_0) \in L^2([0, T] \times \Omega; H^1) \cap L^2(\Omega; C(0, T; H^0))$ and there exists a constant C , depending only on K and T , such that*

$$(3.3) \quad \sup_{0 \leq t \leq T} E\|q^{\mathcal{R}}(t, q_0)\|_k^2 \leq CE \left\{ \|q_0\|_k^2 + \int_0^T [\|f(s, \mu)\|_k^2 + \|g\|_{k+1}^2] ds \right\}, \quad k = 0, 1.$$

Moreover, for any $p \geq 2$, there exists a constant $C(p)$ such that

$$(3.4) \quad \sup_{0 \leq t \leq T} E\|q^{\mathcal{R}}(t, q_0)\|_k^p \leq C(p)E \left\{ \|q_0\|_k^p + \int_0^T [\|f(s, \mu)\|_k^p + \|g\|_{k+1}^p] ds \right\},$$

$k = 0, 1$.

The main idea of the present paper is based on the fact that the first-order differential operator M generates a strongly continuous group on H^0 . The following proposition states the detailed properties of M , with the proof provided in the Appendix.

PROPOSITION 3.2. *On the Hilbert space H^0 , define an operator M by (2.2) with the domain $D(M) := H^1$, then*

(i) *M can be extended to a closed operator (still denoted by M) that generates a strongly continuous group $\{e^{Mt} : -\infty < t < +\infty\}$ on H^0 . Moreover, H^1 is an invariant subspace of e^{Mt} for each t , and there exists a constant N (the same as the constant in (3.2)), such that*

$$(3.5) \quad \|e^{Mt}\|_{L(H^0 \rightarrow H^0)} \leq e^{N|t|},$$

$$(3.6) \quad \|e^{Mt}\|_{L(H^1 \rightarrow H^1)} \leq e^{N|t|}, \quad \text{for any } t \in (-\infty, +\infty);$$

(ii) *Denote by M^* the adjoint operator of M on H^0 , then $H^1 \subset D(M^*)$ and M^* also generates a strongly continuous group $\{e^{M^*t} = (e^{Mt})^* : -\infty < t < +\infty\}$ on H^0 . Moreover, H^1 is an invariant subspace of e^{M^*t} for each t , and with the same constant N , we have*

$$(3.7) \quad \|e^{M^*t}\|_{L(H^0 \rightarrow H^0)} \leq e^{N|t|},$$

$$(3.8) \quad \|e^{M^*t}\|_{L(H^1 \rightarrow H^1)} \leq e^{N|t|}, \quad \text{for any } t \in (-\infty, +\infty);$$

(iii) *Define two operators M^2, M^{*2} from H^1 to H^{-1} by the following formula:*

$$(3.9) \quad \langle M^2 \phi, \psi \rangle_0 = \langle M\phi, M^* \psi \rangle_0 = \langle \phi, M^{*2} \psi \rangle_0, \quad \text{for } \phi, \psi \in H^1,$$

*then M^2 and M^{*2} are bounded linear operators from H^1 to H^{-1} .*

From now on, when we write M, M^*, M^2 , and M^{*2} , it is always understood to be in the sense of that in Proposition 3.2.

The following lemma will play an essential role in this paper.

LEMMA 3.2. *Let D be a set in R^d such that*

$$(3.10) \quad D \text{ is bounded, open, and with smooth boundary.}$$

Define $W_D[0, T] := \{\phi : \phi \in L^2(0, T; H^1(D)), d\phi/dt \in L^2(0, T; H^{-1}(D))\}$ with the norm

$$(3.11) \quad \|\phi\|_{W_D[0, T]} := \left(\int_0^T \|\phi(t)\|_{1,D}^2 dt + \int_0^T \|d\phi(t)/dt\|_{-1,D}^2 dt \right)^{1/2},$$

where $H^k(D)$ is the Sobolev space $W_2^k(D)$ with the Sobolev norm $\|\cdot\|_{k,D}$. Then the embedding $W_D[0, T] \rightarrow L^2(0, T; H^0(D))$ is compact.

Proof. Since the embedding $H^1(D) \rightarrow H^0(D)$ is compact, the result follows from [11, Thm. 5.1, p. 58]. \square

PROPOSITION 3.3. *Assume (A1)–(A3), $q_0 \in H^1$, and $\mathcal{R} \in \mathbb{R}$. Let $q(\cdot) = q^{\mathcal{R}}(\cdot, q_0)$ be the solution of (2.9). Set $p(t) := e^{-MW(t)} q(t)$, then p satisfies*

$$(3.12) \quad \begin{aligned} d(p(t), \phi)_0 &= \{ \langle (A(t, \mu) - \frac{1}{2}M^2)q(t), e^{-M^*W(t)} \phi \rangle_0 \\ &\quad + (e^{-MW(t)}(f(t, \mu) - Mg), \phi)_0 \} dt + (e^{-MW(t)}g, \phi)_0 dW(t), \end{aligned}$$

for any $\phi \in H^1$.

Proof. Take $\phi, \psi \in H^1$. Define $\rho(t) := (e^{-M^*t} \phi, \psi)_0$. By Proposition 3.2,

$$d\rho(t)/dt = -(M^* e^{-M^*t} \phi, \psi)_0 = -(e^{-M^*t} \phi, M\psi)_0,$$

$$d^2\rho(t)/dt^2 = (M^* e^{-M^*t} \phi, M\psi)_0 = \langle M^{*2} e^{-M^*t} \phi, \psi \rangle_0.$$

Hence by Itô's formula, we have

$$(3.13) \quad d\rho(W(t)) = \frac{1}{2} \langle M^{*2} e^{-M^*W(t)} \phi, \psi \rangle_0 dt - \langle M^* e^{-M^*W(t)} \phi, \psi \rangle_0 dW(t);$$

namely, we have the following formula in the space H^{-1} :

$$(3.14) \quad d(e^{-M^*W(t)} \phi) = \frac{1}{2} M^{*2} e^{-M^*W(t)} \phi dt - M^* e^{-M^*W(t)} \phi dW(t).$$

Therefore, again by Itô's formula,

$$\begin{aligned}
 d(p(t), \phi)_0 &= d(q(t), e^{-M^* W(t)} \phi)_0 \\
 &= \langle A(t, \mu)q(t) + f(t, \mu), e^{-M^* W(t)} \phi \rangle_0 dt \\
 &\quad + \langle Mq(t) + g, e^{-M^* W(t)} \phi \rangle_0 dW(t) \\
 &\quad + \langle q(t), \frac{1}{2} M^{*2} e^{-M^* W(t)} \phi \rangle_0 dt - \langle q(t), M^* e^{-M^* W(t)} \phi \rangle_0 dW(t) \\
 &\quad - \langle Mq(t) + g, M^* e^{-M^* W(t)} \phi \rangle_0 dt \\
 &= \{ \langle (A(t, \mu) - \frac{1}{2} M^2)q(t), e^{-M^* W(t)} \phi \rangle_0 \\
 &\quad + \langle e^{-M^* W(t)} (f(t, \mu) - Mg), \phi \rangle_0 \} dt \\
 &\quad + \langle e^{-M^* W(t)} g, \phi \rangle_0 dW(t).
 \end{aligned}
 \tag{3.15}$$

This proves (3.12). \square

We need an additional lemma for technical reasons.

LEMMA 3.3. *For some $r > 0$, we assume (A1)–(A3), (A3)_r, and $q_0 \in H^1 \cap L_r^2$. For $\mathcal{R} \in \mathcal{R}$, let $q(t, x) = q^{\mathcal{R}}(t, x, q_0)$ be the solution of (2.9), then there exists a constant C , independent of \mathcal{R} , such that*

$$E \int_{|x| > \rho} |q(t, x)|^2 dx \leq C/(1 + \rho^2)^r, \quad \text{for any } \rho > 0.
 \tag{3.16}$$

Proof. By Krylov and Rozovskii [10], we have a constant C' such that

$$\sup_{0 \leq t \leq T} E \|q(t)\|_{0,r}^2 \leq C' E \left\{ \|q_0\|_{0,r}^2 + \int_0^T (\|f(t, \mu)\|_{0,r}^2 + \|g\|_{1,r}^2) dr \right\} \leq C;$$

hence

$$E \int_{|x| > \rho} |q(t, x)|^2 dx \leq E \|q(t)\|_{0,r}^2 / (1 + \rho^2)^r \leq C/(1 + \rho^2)^r. \quad \square$$

THEOREM 3.1. *Assume (A1)–(A3), (A3)_r, and $q_0 \in H^1 \cap L_r^2$ for some $r > 0$. Denote by $q^{\mathcal{R}}(\cdot, q_0)$ the solution of (2.9) for $\mathcal{R} \in \mathcal{R}$. If $\mathcal{R}_n \rightarrow \mathcal{R}$, then*

$$q^{\mathcal{R}_n}(\cdot, q_0) \rightarrow q^{\mathcal{R}}(\cdot, q_0) \text{ in law, as } L^2(0, T; H^0)\text{-}r.v.;
 \tag{3.17}$$

$$q^{\mathcal{R}_n}(T, q_0) \rightarrow q^{\mathcal{R}}(T, q_0) \text{ in law, as } H^0\text{-}r.v.
 \tag{3.18}$$

Proof. Suppose $\mathcal{R}_n = (W_n, \mu_n)$ and $\mathcal{R} = (W, \mu)$. We write $q_n(\cdot) := q^{\mathcal{R}_n}(\cdot, q_0)$ for simplicity. Define $p_n(t) := e^{-M^* W_n(t)} q_n(t)$, $p_{n,1}(t) := \int_0^t e^{-M^* W_n(s)} g dW_n(s)$, and $p_{n,2}(t) := p_n(t) - p_{n,1}(t)$. Then, by Propositions 3.1 and 3.2, we have

$$\begin{aligned}
 E \int_0^T \|p_n(t)\|_1^2 dt &\leq E \int_0^T e^{2N|W_n(t)|} \|q_n(t)\|_1^2 dt \\
 &\leq E \left(\sup_{0 \leq t \leq T} e^{2N|W_n(t)|} \int_0^T \|q_n(t)\|_1^2 dt \right) \\
 &\leq \left(E \sup_{0 \leq t \leq T} e^{4N|W_n(t)|} \right)^{1/2} \left(E \int_0^T \|q_n(t)\|_1^4 dt \right)^{1/2} \cdot T^{1/2}.
 \end{aligned}
 \tag{3.19}$$

It is known from Fernique's lemma [5] that $\sup_n (E \sup_{0 \leq t \leq T} e^{4N|W_n(t)|}) < +\infty$; hence (3.19) yields

$$\sup_n E \int_0^T \|p_n(t)\|_1^2 dt < +\infty.
 \tag{3.20}$$

Similarly, we have

$$(3.21) \quad \sup_n E \int_0^T \|p_{n,1}(t)\|_1^2 dt \leq \sup_n E \int_0^T \int_0^t e^{2N|W_n(s)|} \|g\|_1^2 ds dt < +\infty.$$

Combining (3.20) and (3.21) gives

$$(3.22) \quad \sup_n E \int_0^T \|p_{n,2}(t)\|_1^2 dt < +\infty.$$

On the other hand, by virtue of Proposition 3.3,

$$(3.23) \quad \begin{aligned} d(p_{n,2}(t), \phi)_0 &= \{ \langle (A(t, \mu_n) - \frac{1}{2}M^2)q_n(t), e^{-M^*W_n(t)}\phi \rangle_0 \\ &\quad + \langle e^{-M^*W_n(t)}(f(t, \mu_n) - Mg), \phi \rangle_0 \} dt, \end{aligned}$$

for any $\phi \in H^1$.

Hence

$$\begin{aligned} |\langle dp_{n,2}(t)/dt, \phi \rangle_0| &\leq \| (A(t, \mu_n) - \frac{1}{2}M^2)q_n(t) \|_{-1} \| e^{-M^*W_n(t)}\phi \|_1 \\ &\quad + e^{N|W_n(t)|} \| f(t, \mu_n) - Mg \|_0 \| \phi \|_0 \\ &\leq \text{const } e^{N|W_n(t)|} (\|q_n(t)\|_1 + 1) \| \phi \|_1, \end{aligned}$$

for any $\phi \in H^1$.

This yields

$$(3.24) \quad \begin{aligned} \sup_n E \int_0^T \|dp_{n,2}(t)/dt\|_{-1}^2 dt \\ \leq \text{const } \sup_n E \int_0^T e^{2N|W_n(t)|} (\|q_n(t)\|_1^2 + 1) dt \\ < +\infty. \end{aligned}$$

Equations (3.22) and (3.24) imply that there exists a constant C_1 that is independent of n such that, for any $D \subset R^d$ with the property (3.10),

$$(3.25) \quad E \|p_{n,2}\|_{W_D[0,T]}^2 \leq C_1.$$

Let $D_k := \{x \in R^d : |x| < k\}$ for $k = 1, 2, \dots$. Define a metric \bar{d} on $L^2(0, T; H^0)$ by

$$(3.26) \quad \bar{d}(\phi, \psi) := \sum_{k=1}^{\infty} \frac{1}{2^k} \min \left\{ 1, \left(\int_0^T \|\phi(t) - \psi(t)\|_{0,D_k}^2 dt \right)^{1/2} \right\}.$$

We denote by $\bar{L}^2(0, T; H^0)$ the completion of $L^2(0, T; H^0)$ by \bar{d} . For $\lambda > 0$,

$$B_\lambda := \{ \phi \in \bar{L}^2(0, T; H^0) : \|\phi\|_{W_{D_k}[0,T]} \leq (2^k \lambda)^{1/2}, \quad k = 1, 2, \dots \}$$

is compact in $\bar{L}^2(0, T; H^0)$ due to Lemma 3.2. Now (3.25) yields

$$(3.27) \quad P(p_{n,2} \in B_\lambda) \leq \sum_{k=1}^{\infty} \frac{1}{2^k \lambda} C_1 \leq C_1 / \lambda, \quad \text{for any } \lambda > 0;$$

hence, $\{p_{n,2}\}$ is tight as $\bar{L}^2(0, T; H^0)$ -r.v. (cf. [7, Def. 2.2, p. 7]). Noting the compactness of Λ , $\{(W_n, \mu_n, p_{n,2})\}$ is tight as $C(0, T; R^1) \times \Lambda \times \bar{L}^2(0, T; H^0)$ -r.v. Hence by Skorohod's theorem, we can choose a subsequence $\{n'\}$ and have $(\hat{W}_{n'}, \hat{\mu}_{n'}, \hat{p}_{n',2})$, $(\hat{W}, \hat{\mu}, \hat{p}_2)$ on a suitable probability space $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{P})$, such that

$$(3.28) \quad \text{law of } (\hat{W}_{n'}, \hat{\mu}_{n'}, \hat{p}_{n',2}) = \text{law of } (W_{n'}, \mu_{n'}, p_{n',2}),$$

and \hat{P} -almost surely

$$(3.29) \quad \hat{W}_{n'} \rightarrow \hat{W} \quad \text{uniformly on } [0, T],$$

$$(3.30) \quad \hat{\mu}_{n'} \rightarrow \hat{\mu} \quad \text{weakly on } \Lambda,$$

$$(3.31) \quad \hat{p}_{n',2} \rightarrow \hat{p}_2 \quad \text{in } \bar{L}^2(0, T; H^0), \text{ as } n' \rightarrow +\infty.$$

Define

$$\hat{p}_{n',1}(t) := \int_0^t e^{-M\hat{W}_{n'}(s)} g d\hat{W}_{n'}(s), \quad \hat{p}_1(t) := \int_0^t e^{-M\hat{W}(s)} g d\hat{W}(s).$$

By virtue of (3.29) and the strong continuity of the group $\{e^{Mt}\}$, we have for fixed $(s, \hat{\omega})$, $\|e^{-M\hat{W}_{n'}(s)} g - e^{-M\hat{W}(s)} g\|_0^2 \rightarrow 0$ as $n' \rightarrow \infty$. By Fernique's lemma, it is easy to check that

$$\sup_{n'} \hat{E} \int_0^T \|e^{-M\hat{W}_{n'}(s)} g - e^{-M\hat{W}(s)} g\|_0^4 ds < +\infty,$$

which means that $\{\|e^{-M\hat{W}_{n'}(s)} g - e^{-M\hat{W}(s)} g\|_0^2\}$ is uniformly integrable on $[0, T] \times \hat{\Omega}$; hence

$$(3.32) \quad e^{-M\hat{W}_{n'}(s)} g \rightarrow e^{-M\hat{W}(s)} g \quad \text{in } L^2([0, T] \times \hat{\Omega}; H^0), \text{ as } n' \rightarrow \infty.$$

Combining (3.32) with (3.29), by a similar argument to that in [12, Lemma 3.3], we get

$$(3.33) \quad \hat{E} \int_0^T \|\hat{p}_{n',1}(t) - \hat{p}_1(t)\|_0^2 dt \rightarrow 0, \quad \text{as } n' \rightarrow \infty.$$

Define

$$\hat{p}_{n'}(t) := \hat{p}_{n',1}(t) + \hat{p}_{n',2}(t), \quad \hat{p}(t) := \hat{p}_1(t) + \hat{p}_2(t);$$

then (3.31) and (3.33) yields and there exists a subsequence of $\{n'\}$ (still denoted by $\{n'\}$) such that

$$(3.34) \quad \hat{p}_{n'} \rightarrow \hat{p} \quad \text{in } \bar{L}^2(0, T; H^0), \quad \text{as } n' \rightarrow \infty, \quad \hat{P}\text{-a.s.}$$

Define

$$\hat{q}_{n'}(t) := e^{M\hat{W}_{n'}(t)} \hat{p}_{n'}(t), \quad \hat{q}(t) := e^{M\hat{W}(t)} \hat{p}(t).$$

Observing (3.28), we have

$$(3.35) \quad \text{law of } (\hat{W}_{n'}, \hat{\mu}_{n'}, \hat{q}_{n'}) = \text{law of } (W_{n'}, \mu_{n'}, q_{n'}).$$

Now we want to show

$$(3.36) \quad \hat{E} \int_0^T \|\hat{q}_{n'}(t) - \hat{q}(t)\|_0^2 dt \rightarrow 0, \quad \text{as } n' \rightarrow +\infty.$$

Indeed, for any $D \subset R^d$ that satisfies (3.10), we have

$$\begin{aligned} & \hat{E} \int_0^T \|\hat{q}_{n'}(t) - \hat{q}(t)\|_{0,D}^2 dt \\ &= \hat{E} \int_0^T \|e^{M\hat{W}_{n'}(t)} \hat{p}_{n'}(t) - e^{M\hat{W}(t)} \hat{p}(t)\|_{0,D}^2 dt \\ &\leq 2\hat{E} \int_0^T \|e^{M\hat{W}_{n'}(t)} (\hat{p}_{n'}(t) - \hat{p}(t))\|_{0,D}^2 dt \\ &\quad + 2\hat{E} \int_0^T \|(e^{M\hat{W}_{n'}(t)} - e^{M\hat{W}(t)}) \hat{p}(t)\|_{0,D}^2 dt = I_3 + I_4, \quad \text{say,} \end{aligned} \quad (3.37)$$

$$\begin{aligned}
I_3 &\leq 2\hat{E} \int_0^T e^{2N|\hat{W}_{n'}(t)|} \|\hat{p}_{n'}(t) - \hat{p}(t)\|_{0,D}^2 dt \\
&\leq 2\hat{E} \left(\sup_{0 \leq t \leq T} e^{2N|\hat{W}_{n'}(t)|} \int_0^T \|\hat{p}_{n'}(t) - \hat{p}(t)\|_{0,D}^2 dt \right) \\
&\leq \text{const} \left\{ \hat{E} \left(\int_0^T \|\hat{p}_{n'}(t) - \hat{p}(t)\|_{0,D}^2 dt \right)^2 \right\}^{1/2}.
\end{aligned}$$

By (3.4), it is easy to see that $\{(\int_0^T \|\hat{p}_{n'}(t) - \hat{p}(t)\|_{0,D}^2 dt)^2\}$ is uniformly integrable on $(\hat{\Omega}, \hat{\mathcal{F}}, \hat{P})$. Thus (3.34) derives $I_3 \rightarrow 0$ as $n' \rightarrow \infty$. Moreover, since for fixed $(t, \hat{\omega})$, $\|(e^{M\hat{W}_{n'}(t)} - e^{M\hat{W}(t)})\hat{p}(t)\|_{0,D}^2 \rightarrow 0$ as $n' \rightarrow \infty$, so by the uniformly integrability of $\{(e^{M\hat{W}_{n'}(t)} - e^{M\hat{W}(t)})\hat{p}(t)\|_{0,D}^2\}$ on $[0, T] \times \hat{\Omega}$, we have $I_4 \rightarrow 0$, as $n' \rightarrow +\infty$. Now we arrive at

$$(3.38) \quad \hat{E} \|\hat{q}_{n'} - \hat{q}\|_{L^2(0,T;H^0)} \rightarrow 0, \quad \text{as } n' \rightarrow +\infty.$$

By (3.38) and (3.16), we can write

$$\begin{aligned}
(3.39) \quad \hat{E} \int_0^T \int_{|x|>\rho} |\hat{q}(t, x)|^2 dx dt &= \lim_{k \rightarrow \infty} \lim_{n' \rightarrow \infty} \hat{E} \int_0^T \int_{\rho < |x| < k} |\hat{q}_{n'}(t, x)|^2 dx dt \\
&\leq CT/(1+\rho^2)^r \rightarrow 0, \quad \text{as } \rho \rightarrow +\infty.
\end{aligned}$$

Hence $\hat{q} \in L^2([0, T] \times \hat{\Omega}; H^0)$, and (3.36) follows from (3.38), (3.39), and (3.16).

By virtue of (3.38), we can prove that \hat{q} is just the response for $\hat{\mathcal{R}} := (\hat{W}, \hat{\mu})$ by a standard argument (cf. [1], [12], and [13]). Noting (3.28)–(3.36), we have proved (3.17). For (3.18), we can apply entirely the same argument, observing the compactness of the embedding: $H^1(D) \rightarrow H^0(D)$. The proof is now complete. \square

Remark 3.1. Condition (A3)_r and $q_0 \in L_r^2$ are required to ensure (3.36) from (3.38). If we drop these conditions, (3.38) still holds, and therefore \hat{q} is still the response of $\hat{\mathcal{R}}$ by the above proof. Thus we have the following corollary.

COROLLARY 3.2. *If we only assume (A1)–(A3) and $q_0 \in H^1$, then whenever $\mathcal{R}_n \rightarrow \mathcal{R}$, we have*

$$(3.40) \quad q^{\mathcal{R}_n}(\cdot, q_0) \rightarrow q^{\mathcal{R}}(\cdot, q_0) \text{ in law, as } w - L^2(0, T; H^1)\text{-r.v.};$$

$$(3.41) \quad q^{\mathcal{R}_n}(T, q_0) \rightarrow q^{\mathcal{R}}(T, q_0) \text{ in law, as } w - H^1\text{-r.v.,}$$

where “ $w - X$ ” denotes the space X endowed with the weak topology.

Now we are in the position to prove the existence of optimal relaxed control for system (2.9) with the cost functional (2.12).

THEOREM 3.2. *In addition to the same assumptions as in Theorem 3.1, we assume*

$$(3.42) \quad F, G \text{ are continuous mappings from } L^2(0, T; H^0) \text{ and } H^0 \text{ to } R^1, \text{ respectively, and there exists } K > 0, \text{ such that}$$

$$|F(\phi)| \leq K(1 + \|\phi\|_{L^2(0,T;H^0)}); \quad |G(\psi)| \leq K(1 + \|\psi\|_0).$$

Then there exists an optimal relaxed control for the system (2.9) with the cost functional (2.12).

Proof. By Theorem 3.1, $J(q_0, \cdot)$ is continuous. However, Π is a compact metric space, which yields the existence of an optimal relaxed control. \square

COROLLARY 3.3. *In addition to the assumptions of Corollary 3.2, we assume*

$$(3.43) \quad F, G \text{ are weakly continuous mappings from } L^2(0, T; H^1) \text{ and } H^1 \text{ to } R^1, \text{ respectively, and satisfy linear growth conditions as in (3.42).}$$

Then there exists an optimal relaxed control.

Proof. The result is a direct consequence of Corollary 3.2. \square

4. Nondegenerate case. In this section we consider the optimal relaxed control problem (2.9) and (2.12) under the assumption that $(a^{ij} - \frac{1}{2}\sigma^i\sigma^j)_{ij}$ is uniformly positive definite. It will be proved that the existence of an optimal relaxed control still holds even when $q_0 \in H^0$.

The following assumptions remain in force throughout this section:

(B1) Same as (A1) in § 3;

(B2) $a^{ij} = a^{ji}$, $i, j = 1, 2, \dots, d$, and there exists $\alpha > 0$ such that

$$(a^{ij}(x, u) - \frac{1}{2}\sigma^i(x)\sigma^j(x))\xi_i\xi_j \geq \alpha|\xi|^2, \quad \text{for any } (x, u, \xi) \in R^d \times \Gamma \times R^d;$$

(B3) $f(\cdot, u) \in H^{-1}$, $g \in H^0$; the absolute values of f, g together with the H^{-1} -norm of $f(\cdot, u)$ and the H^0 -norm of g do not exceed K .

The following proposition is an easy variant of the results in Krylov and Rozovskii [8] and Pardoux [14].

PROPOSITION 4.1. *Suppose $q_0 \in H^0$, then for any $\mathcal{R} \in \mathcal{R}$, (2.9) has a unique solution $q^{\mathcal{R}}(\cdot, q_0) \in L^2([0, T] \times \Omega; H^1) \cap L^2(\Omega; C(0, T; H^0))$ and there exists a constant C , depending only on K, T , and α , such that*

$$(4.1) \quad \sup_{0 \leq t \leq T} E \|q^{\mathcal{R}}(t, q_0)\|_0^2 \leq CE \left\{ \|q_0\|_0^2 + \int_0^T [\|f(s, \mu)\|_{-1}^2 + \|g\|_0^2] ds \right\},$$

$$(4.2) \quad E \left(\int_0^T \|q^{\mathcal{R}}(t, q_0)\|_1^2 dt \right)^2 \leq CE \left\{ \|q_0\|_0^4 + \int_0^T [\|f(s, \mu)\|_{-1}^4 + \|g\|_0^4] ds \right\}.$$

Moreover, for any $p \geq 2$, there exists a constant $C(p)$ such that

$$(4.3) \quad \sup_{0 \leq t \leq T} E \|q^{\mathcal{R}}(t, q_0)\|_0^p \leq C(p)E \left\{ \|q_0\|_0^p + \int_0^T [\|f(s, \mu)\|_{-1}^p + \|g\|_0^p] ds \right\}.$$

THEOREM 4.1. *Assume $q_0 \in H^0$. Denote by $q^{\mathcal{R}}(\cdot, q_0)$ the response for $\mathcal{R} = (W, \mu)$ of the system. Then whenever $\mathcal{R}_n \rightarrow \mathcal{R}$, we have*

$$(4.4) \quad q^{\mathcal{R}_n}(\cdot, q_0) \rightarrow q^{\mathcal{R}}(\cdot, q_0) \text{ in law, as } w - L^2(0, T; H^1)\text{-r.v.};$$

$$(4.5) \quad q^{\mathcal{R}_n}(T, q_0) \rightarrow q^{\mathcal{R}}(T, q_0) \text{ in law, as } w - H^0\text{-r.v.}$$

Proof. For simplicity, we denote $q_n(\cdot) := q^{\mathcal{R}_n}(\cdot, q_0)$, $q(\cdot) := q^{\mathcal{R}}(\cdot, q_0)$, for $\mathcal{R}_n = (W_n, \mu_n)$, $\mathcal{R} = (W, \mu)$.

Define $p_n(t) := e^{-MW_n(t)} q_n(t)$, then

$$\begin{aligned} E \int_0^T \|p_n(t)\|_1^2 dt &\leq E \int_0^T e^{2N|W_n(t)|} \|q_n(t)\|_1^2 dt \\ (4.6) \quad &\leq E \left(\sup_{0 \leq t \leq T} e^{2N|W_n(t)|} \int_0^T \|q_n(t)\|_1^2 dt \right) \\ &\leq \left(E \sup_{0 \leq t \leq T} e^{4N|W_n(t)|} \right)^{1/2} \left\{ E \left(\int_0^T \|q_n(t)\|_1^2 dt \right)^2 \right\}^{1/2}. \end{aligned}$$

So by (4.2), $\sup_n E \int_0^T \|p_n(t)\|_1^2 dt < +\infty$. Now by applying entirely the same argument as that in the proof of Theorem 3.1 and Corollary 3.2, we obtain (4.4). Noting (4.1), we also get (4.5). (But we no longer have $q_n(T) \rightarrow q(T)$ as $w - H^1$ -random variable.) The proof is complete. \square

Now we establish the existence of an optimal relaxed control for the system (2.9) with the cost functional (2.12).

THEOREM 4.2. *Suppose $q_0 \in H^0$ and F, G are weakly continuous mappings from $L^2(0, T; H^1)$ and H^0 , respectively, to R^1 , and they satisfy linear growth conditions. Then there exists an optimal relaxed control.*

5. Discussions and applications.

5.1. Multidimensional Wiener process cases. The method employed in the previous sections is somewhat similar to the *time change* technique in stochastic analysis. This method, however, fails to be effective in general for the system as follows:

$$(5.1) \quad \begin{aligned} dq(t) &= (A(t, \mu)q(t) + f(t, \mu)) dt + \sum_{k=1}^{d'} (M_k q(t) + g_k) dW^k(t), \\ q(0) &= q_0, \end{aligned}$$

where $W := (W^1, \dots, W^{d'})$ is a d' -dimensional Wiener process, and A, M_k have the same forms as (2.1) and (2.2).

However, in some special cases, we can still treat (5.1) by a similar argument to the one-dimensional Wiener process cases.

THEOREM 5.1. *Assume the coefficients of A, M_k ($k = 1, 2, \dots, d'$) satisfy (A1) and (A2) of § 3 (respectively, (B1) and (B2) of § 4). Assume further that*

$$(5.2) \quad M_k M_j = M_j M_k, \quad \text{for } k \neq j,$$

then Theorem 3.2 and Corollary 3.3 (respectively, Theorem 4.2) hold for the system (5.1).

Proof. We will show this for $d' = 2$ for simplicity. By the proofs in the previous sections, it suffices to prove that we can construct a transformation $q \rightarrow p$, such that p satisfies an SPDE whose diffusion term is independent of the state (refer to Proposition 3.3). To this end, define $p_1(t) := e^{-M_1 W^1(t)} q(t)$, then a similar calculation to that of Proposition 3.3 gives that p_1 satisfies the following differential formula in H^{-1} space:

$$(5.3) \quad \begin{aligned} dp_1(t) &= e^{-M_1 W^1(t)} \{ (A(t, \mu) - \tfrac{1}{2} M_1^2) q(t) + f(t, \mu) - M_1 g_1 \} dt \\ &\quad + e^{-M_1 W^1(t)} g_1 dW^1(t) + \{ e^{-M_1(t) W^1(t)} (M_2 q(t) + g_2) \} dW^2(t). \end{aligned}$$

Put $p(t) := e^{-M_2 W^2(t)} p_1(t)$, then, for any $\phi \in H^1$,

$$(5.4) \quad \begin{aligned} d(p(t), \phi)_0 &= d(p_1(t), e^{-M_2^* W^2(t)} \phi)_0 \\ &= \langle e^{-M_1 W^1(t)} \{ (A(t, \mu) - \tfrac{1}{2} M_1^2) q(t) + f(t, \mu) - M_1 g_1 \}, e^{-M_2^* W^2(t)} \phi \rangle_0 dt \\ &\quad + (e^{-M_1 W^1(t)} g_1, e^{-M_2^* W^2(t)} \phi)_0 dW^1(t) \\ &\quad + (e^{-M_1 W^1(t)} (M_2 q(t) + g_2), e^{-M_2^* W^2(t)} \phi)_0 dW^2(t) \\ &\quad + (p_1(t), \tfrac{1}{2} M_2^{*2} e^{-M_2^* W^2(t)} \phi)_0 dt - (p_1(t), M_2^* e^{-M_2^* W^2(t)} \phi)_0 dW^2(t) \\ &\quad + (e^{-M_1 W^1(t)} (M_2 q(t) + g_2), -M_2^* e^{-M_2^* W^2(t)} \phi)_0 dt \\ &= \langle (A(t, \mu) - \tfrac{1}{2} M_1^2 - \tfrac{1}{2} M_2^2) q(t), e^{-M_1 W^1(t)} e^{-M_2^* W^2(t)} \phi \rangle_0 dt \\ &\quad + (e^{-M_2 W^2(t)} e^{-M_1 W^1(t)} (f(t, \mu) - M_1 g_1 - M_2 g_2), \phi)_0 dt \\ &\quad + (e^{-M_2 W^2(t)} e^{-M_1 W^1(t)} g_1, \phi)_0 dW^1(t) \\ &\quad + (e^{-M_2 W^2(t)} e^{-M_1 W^1(t)} g_2, \phi)_0 dW^2(t). \end{aligned}$$

Note that in the above calculation, we have repeatedly employed the fact that $e^{M_1 t'} M_2|_{H^1} = M_2 e^{M_1 t'}|_{H^1}$, which is an easy consequence of the fact that $M_1 M_2 = M_2 M_1$. Now (5.4) has a similar form to (3.12), which allows us to apply the same argument as those in §§ 3 and 4 to obtain the desired results. \square

Remark 5.1. Equation (5.2) holds when the coefficients of $\{M_k\}$ are constants or satisfy some linear dependence conditions.

Remark 5.2. In particular, (5.2) holds for the following kind of systems:

$$(5.5) \quad \begin{aligned} dq(t) &= (A(t, \mu)q(t) + f(t, \mu)) dt + \sum_{k=1}^{d'} (h_k q(t) + g_k) dW^k(t), \\ q(0) &= q_0, \end{aligned}$$

where $h_k: R^d \rightarrow R^1$. The above system has been studied by Bensoussan and Nisio [1] (with $f = g_k = 0$). Their result is [1, Thm. 5.2]: if (a^{ij}) is uniformly positive definite and $q_0 \in H^2(R^d)$, then there exists an optimal relaxed control. Now we know the condition “ $q_0 \in H^2(R^d)$ ” can be weakened to “ $q_0 \in L^2(R^d)$ ”; on the other hand, if we assume $q_0 \in H^1(R^d)$, then the result is valid even if (a^{ij}) is degenerate.

5.2. Stochastic control with partial observation. First, let us remark that the results in §§ 3 and 4 are easily extended to the following kind of system:

$$(5.6) \quad \begin{aligned} dq(t) &= (\tilde{A}(t, W(t), \mu)q(t) + \tilde{f}(t, W(t), \mu)) dt + (Mq(t) + g(W(t))) dW(t), \\ q(0) &= q_0, \end{aligned}$$

where

$$\begin{aligned} \tilde{A}(t, w, \mu)\phi(x) &:= \partial_i(\tilde{a}^{ij}(t, x, w, \mu)\partial_j\phi(x)) + \tilde{b}^i(t, x, w, \mu)\partial_i\phi(x) \\ &\quad + \tilde{c}(t, x, w, \mu)\phi(x), \\ M\phi(x) &:= \sigma^i(x)\partial_i\phi(x) + h(x)\phi(x), \\ \tilde{a}^{ij}(t, x, w, \mu) &:= \int_{\Gamma} a^{ij}(x, w, u)\mu'(t, du), \quad \text{etc.} \end{aligned}$$

Let B and \hat{B} be independent Wiener processes on a probability space (Ω, \mathcal{F}, P) , with values in R^1 and R^d , respectively. Consider the following SPDE in R^d :

$$(5.7) \quad \begin{aligned} dX(t) &= \gamma(X(t), Y(t), U(t)) dt + \alpha(X(t), Y(t), U(t)) d\hat{B}(t) + \sigma(X(t)) dB(t), \\ X(0) &= \xi, \end{aligned}$$

with the observation

$$(5.8) \quad dY(t) = \kappa(X(t)) dt + dB(t), \quad Y(0) = 0,$$

where U is an admissible control. Note that σ is the correlation between the state and the observation.

Let Ξ and $L: R^d \times R^1 \rightarrow R^1$. The problem is to minimize the cost functional defined by

$$(5.9) \quad J(U) := E \left\{ \int_0^T \Xi(X(t), Y(t)) dt + L(X(t), Y(t)) \right\},$$

over the totality of admissible controls.

By the well-known relationship between the control problem for partially observed diffusions and that for SPDEs (cf. [12]–[15]), we may interpret the problem (5.7)–(5.9) to the one we have treated in the previous sections. Thus we have the following theorem, appealing to Corollary 3.3 and Theorem 4.2.

THEOREM 5.2. *We make the following assumptions:*

- (C1) $\alpha: R^d \times R^1 \times \Gamma \rightarrow R^{d \times d}$, $\sigma: R^d \rightarrow R^d$, $\gamma: R^d \times R^1 \times \Gamma \rightarrow R^d$, $\kappa: R^d \rightarrow R^1$ are continuous functions. α , σ and their derivatives in x up to fourth order do not exceed a constant K in absolute value; γ , κ , and their derivatives in x up to third order do not exceed K in absolute value;

(C2) $\Xi, L: R^d \times R^1 \rightarrow R^1$ are Lipschitz continuous and bounded by K .

Then there exists an optimal relaxed control of the problem (5.7)–(5.9), if we assume, in addition, either of the following two conditions:

- (i) ξ has a density $q_0 \in H^1$;
- (ii) ξ has a density $q_0 \in H^0$, and $\alpha\alpha^*(x, y, u)$ is uniformly positive definite.

Remark 5.3. In [12] and [13], in addition to the higher regularity condition on q_0 , it is required that $(\alpha\alpha^* - 2\sigma\sigma^*)$ be nonnegative definite, which means the correlation cannot be “too large.” In this paper, this restriction is removed.

Let us conclude the paper by two remarks.

Remark 5.4. The relaxed controls turn out to be (usual) admissible controls when assuming some convex conditions (Roxin’s conditions) on the coefficients a^{ij} , etc. The reader may refer to [1], [13] for details.

Remark 5.5. The aim of this paper is to reduce the regularity on the initial state. In the viewpoint of filtering problems, it is more natural to consider the Dirac initial state, although it seems to be a rather difficult problem since the SPDE theory itself with the Dirac initial condition has not been well established.

Appendix. In this appendix, we prove Proposition 3.2.

On the space H^0 , consider the following densely defined operator:

$$(A.1) \quad M: \begin{cases} D(M) & (= \text{the domain of } M) := H^1 \\ M\phi(x) := \sigma^i(x)\partial_i\phi(x) + h(x)\phi(x), & \text{for } \phi \in H^1, \quad x \in R^d. \end{cases}$$

M thus defined is not a closed operator, but it is clearly closable. Denote by \bar{M} the closed extension of M (i.e., the graph of \bar{M} is the closure of the graph of M). On the other hand, M^* , the adjoint operator of M on H^0 , is given by

$$(A.2) \quad M^*: \begin{cases} D(M^*) = \{\phi \in H^0: -\partial_i(\sigma^i\phi) + h\phi \in H^0\}, \\ M^*\phi(x) = -\partial_i(\sigma^i(x)\phi(x)) + h(x)\phi(x), & \text{for } \phi \in D(M^*), \quad x \in R^d. \end{cases}$$

Since H^0 is reflexive, so $\bar{M} \equiv M^{**}$, which is given by

$$(A.3) \quad \bar{M}: \begin{cases} D(\bar{M}) = \{\phi \in H^0: \sigma^i\partial_i\phi + h\phi \in H^0\}, \\ \bar{M}\phi(x) = \sigma^i(x)\partial_i\phi(x) + h(x)\phi(x), & \text{for } \phi \in D(\bar{M}), \quad x \in R^d. \end{cases}$$

Moreover, $\bar{M}^* = M^{***} = M^*$.

LEMMA A.1. There exists a constant N , which is given in Corollary 3.1, such that

$$(A.4) \quad |(\bar{M}\phi, \phi)_0| \leq N \|\phi\|_0^2, \quad \text{for } \phi \in D(\bar{M});$$

$$(A.5) \quad |(\bar{M}^*\phi, \phi)_0| \leq N \|\phi\|_0^2, \quad \text{for } \phi \in D(\bar{M}^*).$$

Proof. Since \bar{M} is the closed extension of M , for $\phi \in D(\bar{M})$, there exist $\{\phi_n\} \subset H^1$, such that $\phi_n \xrightarrow{H^0} \phi$, $M\phi_n \xrightarrow{H^0} \bar{M}\phi$. Hence (A.4) follows from Corollary 3.1. On the other hand, \bar{M}^* is the closed extension of $K\phi := -\partial_i(\sigma^i\phi) + h\phi$ with $D(K) := H^1$, so (A.5) is proved by a similar argument. \square

Proof of Proposition 3.2. Let $\lambda \in R^1$ such that $|\lambda| > N$. For $\phi \in D(\bar{M})$, set $\psi := (\lambda I - \bar{M})\phi$, then

$$(A.6) \quad |(\psi, \phi)_0| \geq |\lambda| \|\phi\|_0^2 - |(\bar{M}\phi, \phi)_0| \geq (|\lambda| - N) \|\phi\|_0^2,$$

which yields $\lambda I - \bar{M}$ is one-to-one, and $\text{Range}(\lambda I - \bar{M})$ is a closed set in H^0 . However, by virtue of (A.5), $\lambda I - \bar{M}^*$ is also one-to-one; thus $\text{Range}(\lambda I - \bar{M}) = \text{Ker}(\lambda I - \bar{M}^*)^\perp = H^0$. Then, by (A.6), λ is in the resolvent set of \bar{M} , and

$$(A.7) \quad \|(\lambda I - \bar{M})^{-1}\|_{L(H^0 \rightarrow H^0)} \leq (|\lambda| - N)^{-1}.$$

By [2, Cor. 17, p. 628], \bar{M} generates a strongly continuous group $\{e^{\bar{M}t}; -\infty < t < +\infty\}$ on H^0 and (3.5) holds.

On the space H^1 , consider the following densely defined operator M_1 :

$$(A.8) \quad M_1: \begin{cases} D(M_1) := H^2, \\ M_1\phi(x) := \sigma^i(x)\partial_i\phi(x) + h(x)\phi(x), \quad \phi \in H^2, \quad x \in R^d. \end{cases}$$

There is also a closed extension \bar{M}_1 of M_1 . By virtue of Corollary 3.1,

$$|(\bar{M}_1\phi, \phi)_1| \leq N\|\phi\|_1^2, \quad \text{for } \phi \in D(\bar{M}_1).$$

So by a similar argument as above, \bar{M}_1 generates a strongly continuous group $\{e^{\bar{M}_1 t}; -\infty < t < +\infty\}$ on H^1 , and

$$(A.9) \quad \|e^{\bar{M}_1 t}\|_{L(H^1 \rightarrow H^1)} \leq e^{N|t|}, \quad \text{for } t \in (-\infty, +\infty).$$

For $\phi \in H^2 \subset D(\bar{M}_1)$, since obviously $\bar{M} \equiv \bar{M}_1$ on $D(\bar{M}_1)$,

$$(A.10) \quad d(e^{\bar{M}_1 t}\phi)/dt = \bar{M}_1 e^{\bar{M}_1 t}\phi = \bar{M}e^{\bar{M}_1 t}\phi,$$

where “ d/dt ” is in H^1 -topology. Consequently, (A.10) holds on H^0 , which means $e^{\bar{M}_1 t}\phi = e^{\bar{M}t}\phi$ for any $\phi \in H^2$. However, H^2 is dense in H^1 , so $e^{\bar{M}_1 t}|_{H^1} \equiv e^{\bar{M}t}$. Now (3.6) follows from (A.9). This proves (i) of Proposition 3.2.

Note \bar{M}^* is also a closed extension of a first-order differential operator K , so (ii) of Proposition 3.2 is proved in a completely analogous way to (i). Moreover, since H^0 is reflexive, $e^{\bar{M}^* t} \equiv (e^{\bar{M}t})^*$.

Finally, (iii) of Proposition 3.2 is clear. \square

Acknowledgments. The author expresses his sincere thanks to Professor M. Nisio for bringing the author's interest to the present subject, and for her encouragement and discussion during the course of this research. Thanks are also due to the referees for many helpful criticism and comments.

REFERENCES

- [1] A. BENSOUSSAN AND M. NISIO, *Non linear semi-group arising in the control of diffusions with partial observation*, Stochastics, 30 (1990), pp. 1–46.
- [2] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part I*, Interscience, New York, 1958.
- [3] G. DA PRATO, M. IANNELLI, AND L. TUBARO, *Some results on linear stochastic differential equations in Hilbert spaces*, Stochastics, 6 (1982), pp. 105–116.
- [4] N. EL KAROUI, D. HUÛ NGUYEN, AND M. JEABLANC-PICQUÉ, *Existence of an optimal Markovian filter for the control under partial observations*, SIAM J. Control Optim., 26 (1988), pp. 1025–1061.
- [5] X. FERNIQUE, *Intégrabilité des vecteurs gaussiens*, C.R. Acad. Sci. Paris Sér. I Math., 270 (1970), pp. 1698–1699.
- [6] W. H. FLEMING AND M. NISIO, *On stochastic relaxed control for partially observed diffusions*, Nagoya Math. J., 93 (1984), pp. 71–108.
- [7] N. IKEDA AND S. WATANABE, *Stochastic Differential Equations and Diffusions Processes*, 2nd ed., Kodansha/North-Holland, Tokyo, 1989.
- [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; Math. USSR-Izv., 11 (1977), pp. 1267–1284.
- [9] ———, *On characteristics of the degenerate parabolic Itô 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 (1982), pp. 75–95; Russian Math. Surveys, 37 (1982), pp. 81–105.
- [11] J. L. LIONS, *Quelques méthodes de résolution des problèmes aux limites non linéaire*, Dunod, Paris, 1969.
- [12] N. NAGASE, *On the existence of optimal control for controlled stochastic partial differential equations*, Nagoya Math. J., 115 (1989), pp. 73–85.

- [13] N. NAGASE AND M. NISIO, *Optimal controls for stochastic partial differential equations*, SIAM J. Control Optim., 28 (1990), pp. 186–213.
- [14] E. PARDOUX, *Stochastic partial differential equations and filtering of diffusion processes*, Stochastics, 3 (1979), pp. 127–167.
- [15] B. L. ROZOVSKIĬ, *Nonnegative L^1 -solutions of second order stochastic parabolic equations with random coefficients*, Steklov Seminar, Stat. Control of Stoch. Proc., 1984, Transl. Math. Monographs, American Mathematical Society, Providence, RI, 1985, pp. 410–427.

A GAME THEORETIC APPROACH TO \mathcal{H}^∞ CONTROL FOR TIME-VARYING SYSTEMS*

DAVID J. N. LIMEBEER†, BRIAN D. O. ANDERSON‡, PRAMOD P. KHARGONEKAR§,
AND MICHAEL GREEN†

Abstract. A representation formula for all controllers that satisfy an \mathcal{L}^∞ -type constraint is derived for time-varying systems. It is now known that a formula based on two indefinite algebraic Riccati equations may be found for time-invariant systems over an infinite time support (see [J. C. Doyle et al., *IEEE Trans. Automat. Control*, AC-34 (1989), pp. 831–847]; [K. Glover and J. C. Doyle, *Systems Control Lett.*, 11 (1988), pp. 167–172]; [K. Glover et al., *SIAM J. Control Optim.*, 29 (1991), pp. 283–324]; [M. Green et al., *SIAM J. Control Optim.*, 28 (1990), pp. 1350–1371]; [D. J. N. Limebeer et al., in *Proc. IEEE conf. on Decision and Control*, Austin, TX, 1988]; [G. Tadmor, *Math. Control Systems Signal Processing*, 3 (1990), pp. 301–324]). In the time-varying case, two indefinite Riccati differential equations are required. A solution to the design problem exists if these equations have a solution on the optimization interval. The derivation of the representation formula illustrated in this paper makes explicit use of linear quadratic differential game theory and extends the work in [J. C. Doyle et al., *IEEE Trans. Automat. Control*, AC-34 (1989), pp. 831–847] and [G. Tadmor, *Math. Control Systems Signal Processing*, 3 (1990), pp. 301–324]. The game theoretic approach is particularly simple, in that the background mathematics required for the sufficient conditions is little more than standard arguments based on “completing the square.”

Key words. \mathcal{H}^∞ -optimal control, game theory, indefinite Riccati equations, four-block general distance problems, worst-case design

1. Introduction. Since the early 1980s, numerous techniques have been developed for solving \mathcal{H}^∞ -control problems. In addition to their individual interest, further insights have been gained by studying the interplay between these various methodologies, and researchers have now acquired a good understanding of several aspects of the theory. Over the last two years, there has been a flurry of activity that has had a significant impact on both the accessibility of the theoretical ideas and the ease of computation. Once an \mathcal{H}^∞ -norm bound has been decided, and provided that a solution exists, the computational burden associated with finding all \mathcal{H}^∞ controllers is essentially the same as that required in solving a linear quadratic Gaussian regulator problem [9], [10], [11], [13], [20], [21], [27]. \mathcal{H}^∞ problems in which perfect information is assumed may be solved using a single Riccati equation with dimension equal to that of the problem [18], [19], [23], while the output-feedback problem requires the solution of two Riccati equations.

The “two Riccati equation” formula for all stabilizing controllers satisfying a closed loop \mathcal{H}^∞ -norm constraint has been derived in a number of ways. A state-space solution via the “four-block” problem is reported in [11], [21], while an alternative approach reminiscent of classical linear quadratic theory may be found in [9]. A solution based on J -spectral factorization theory is given in [2], [13], while the related notion of conjugation is used in [20]. An interesting by-product of this activity has

* Received by the editors April 19, 1989; accepted for publication (in revised form) December 13, 1990.

† Department of Electrical Engineering, Imperial College, Exhibition Road, London, England.

‡ Department of System Engineering, Australian National University, Canberra, Australia.

§ Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109-2122. This author was supported in part by National Science Foundation grants ECS-8451519 and ECS-9096109, and in part by grants from Honeywell, Boeing, General Electric, Army Research Office, and United States Air Force Office of Scientific Research grants AFOSR-88-0020 and AFOSR-90-0053.

been the discovery of a number of new interconnections. In [19], [28], the authors note a connection between \mathcal{H}^∞ -control and game theory. The interplay between indefinite factorization and game theory, probably first noted by Banker [3], has been rediscovered in the more general setting of \mathcal{H}^∞ control [10], [13], [22]. The connection between risk-sensitive optimal control [10], [30] and game theory, originally discovered by Jacobson in the perfect-information case [17], has also received renewed interest in the wider setting of \mathcal{H}^∞ control [7], [10]. More recently still, Tadmor [27] has given results on the finite-horizon time-varying case using the maximum principle.

The aim of the present paper, which builds on the work given in [9] and [27], is to give a solution to the finite-horizon time-varying \mathcal{H}^∞ -control problems, which makes explicit use of the existing theory of linear quadratic games. This approach offers the advantage of introducing game theoretic intuition to the solution path, and also gives a game theoretic interpretation to the change of variable introduced in [9]. We also show that the main infrastructure of a time-varying \mathcal{H}^∞ -control theory may be developed using classical arguments based on completing the square. This simple solution is possible because of the linear quadratic (LQ) nature of the problem, and, as a final comment, we mention that we have been able to remove all the removable assumptions in [27].

In this paper we will consider the finite-dimensional linear time-varying plant

$$(1.1a) \quad \begin{bmatrix} z \\ y \end{bmatrix} = \begin{bmatrix} \mathfrak{P}_{11} & \mathfrak{P}_{12} \\ \mathfrak{P}_{21} & \mathfrak{P}_{22} \end{bmatrix} \begin{bmatrix} w \\ u \end{bmatrix},$$

which has a time-varying state-space realization

$$(1.1b) \quad \begin{matrix} n \\ p \\ q \end{matrix} \begin{bmatrix} \hat{x} \\ z \\ y \end{bmatrix} (t) = \begin{bmatrix} \overset{n}{A} & \overset{l}{B_1} & \overset{m}{B_2} \\ C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{22} \end{bmatrix} (t) \begin{bmatrix} x \\ w \\ u \end{bmatrix} (t), \quad x(0) = 0.$$

The class of controls of interest is given by $u = \mathfrak{R}y$, where \mathfrak{R} is a linear time-varying (LTV) controller. Eliminating u and y gives rise to the closed loop operator $z = \mathfrak{R}w$, where $\mathfrak{R} = \mathfrak{P}_{11} + \mathfrak{P}_{12}\mathfrak{R}(I - \mathfrak{P}_{22}\mathfrak{R})^{-1}\mathfrak{P}_{21}$. Our goals are (1) to give necessary and sufficient conditions for the existence of LTV controllers such that $\|z\|_2 < \gamma\|w\|_2$ for all $w \neq 0$ and a given γ ($\|\cdot\|_2$ denotes the norm on $\mathcal{L}^2[0, T]$), and (2) to characterize all such controllers when they exist. We assume that the matrices in (1.1) have entries that are continuous functions of time, that D_{12} has full column rank m for all $t \in [0, T]$, and that D_{21} has full row rank q for all $t \in [0, T]$. The game theoretic nature of the problem is immediate: Roughly speaking, the w -player tries to maximize the energy in z , while the controller or u -player simultaneously seeks to minimize it.

Section 2 has a large tutorial component and deals with the time-varying problem in which perfect information is assumed. Section 2.1 establishes the necessary conditions for the existence of a solution to the \mathcal{H}^∞ -control problem via a conjugate point argument. We then show that \mathcal{H}^∞ -control problems with perfect information may be solved if and only if a solution to the Riccati differential equation (RDE) exists on $[0, T]$. Section 2.2 presents a representation formula for all solutions to the time-varying \mathcal{H}^∞ -control problem in the perfect-information case. A brief review of some pertinent properties of adjoint systems is given in § 2.3. Section 2.4 partially reconciles the more familiar time-invariant infinite-horizon \mathcal{H}^∞ -control problem with the time-varying finite-horizon case. This reconciliation calls for a connection between the limiting solution of the RDE and its algebraic counterpart.

Section 3 contains the main results of the paper. We begin with an analysis of problems in which both D_{12} and D_{21} are assumed to be square. A solution is found by calculating a plant inverse, and does not require the solution of any Riccati equations. It is interesting that the plant inverse has an observer structure that is indicative of the solution in the general case. Following that, we treat problems in which either D_{12} or D_{21} is square. Every problem of this type requires the solution of a single RDE. Finally, we treat the case in which neither D_{12} nor D_{21} is square, and we show that in these cases two RDEs are required. We give necessary and sufficient conditions for the existence of solutions and characterize all solutions (when they exist).

2. \mathcal{H}^∞ control and linear quadratic differential games. Numerous variants of the linear quadratic differential game have been studied over the last twenty years. As an example of the extensive literature on the subject, we refer the interested reader to Basar and Olsder [4] and the references therein. The purpose of this section is to review those aspects of the existing game theory literature that are relevant to the solution of the time-varying \mathcal{H}^∞ -control problem, and to study the connections between the two. We begin by considering a system with input vectors $u(t)$ and $w(t)$, dynamical description

$$(2.1) \quad \dot{x}(t) = A(t)x(t) + B_1(t)w(t) + B_2(t)u(t), \quad x(0) = 0,$$

and output,

$$(2.2) \quad z(t) = \begin{bmatrix} C(t)x(t) \\ D(t)u(t) \end{bmatrix}.$$

We will assume that $D(t)$ has full column rank, and that it has been scaled so that

$$(2.3) \quad D'(t)D(t) = I.$$

Each of the matrices in (2.1) and (2.2) is assumed to have entries that are continuous functions of time. In the interests of notational compactness, this time dependence will not always be shown from now on.

With (2.1)–(2.3) given, the \mathcal{H}^∞ -control problem is to find a linear causal control $u(t) = K(x(s), w(s), t)$, $0 \leq s \leq t$, such that $\|\mathcal{T}_{zw}\| = \sup_w \{\|\mathcal{T}_{zw}w\|_2 : w \in \mathcal{L}^2[0, T], \|w\|_2 \leq 1\} < \gamma$ for some given $\gamma > 0$. The operator \mathcal{T}_{zw} maps w to z when the control $u(t) = K(\cdot, \cdot, \cdot)$ is in place. If $z = \mathcal{T}_{zw}w$, then $\|\mathcal{T}_{zw}\| < \gamma$ if and only if

$$(2.4) \quad J(K, w) = \int_0^T (z'z - \gamma^2 w'w) dt \leq -\varepsilon \|w\|_2^2$$

for all $w \in \mathcal{L}^2[0, T]$ and some positive ε .

In the language of game theory, the \mathcal{H}^∞ -control problem is a leader-follower game and requires the control system designer to make the first move and select an admissible $u(t) = K(\cdot, \cdot, \cdot)$ that minimizes $J(K, w)$. A control functional $K(\cdot, \cdot, \cdot)$ is admissible if (1) $u(t) \in \mathcal{L}^2[0, T]$ for every $w(t) \in \mathcal{L}^2[0, T]$, and (2) $x(0) = 0$ and $w(t) \equiv 0 \Rightarrow u(t) \equiv 0$; we denote the set of admissible control functionals \mathcal{C} . After the designer's choice of control strategy has been made public, we assume that nature is malicious and selects that $w(t) \in \mathcal{L}^2[0, T]$, which maximizes (2.4). The \mathcal{H}^∞ -control problem therefore has a solution if and only if

$$(2.5) \quad \min_{K \in \mathcal{C}} \max_{w \in \mathcal{L}^2[0, T]} \{J(K, w)\} \leq -\varepsilon \|w\|_2^2$$

is satisfied. In the notation of (2.5), the w -player maximizes $J(K, w)$ over all $w \in \mathcal{L}^2[0, T]$ after the u -player has announced a choice of $K \in \mathcal{C}$.

Before studying the minimax problem (2.5) in detail, it is interesting to reconcile it with the classical linear quadratic optimal regulator. If we allow γ to increase without bound, then it is necessarily the case that the energy in $w(t)$ decreases to zero, rendering the w -player impotent. The game thus degenerates to the optimal regulator [1] as γ is increased without bound.

Our first result is based on a standard completion of squares argument [4, p. 290], and shows that the \mathcal{H}^∞ -control problem has a solution if a particular RDE has a solution on the time interval $[0, T]$.

THEOREM 2.1. *Suppose that the RDE*

$$(2.6) \quad -\dot{P} = PA + A'P - P(B_2B_2' - \gamma^{-2}B_1B_1')P + C'C, \quad P(T) = 0$$

has a solution on $[0, T]$. Then

$$(2.7) \quad u^* = -B_2'Px,$$

$$(2.8) \quad w^* = \gamma^{-2}B_1'Px$$

result in

$$(2.9) \quad \|z\|_2^2 - \gamma^2\|w\|_2^2 = \|u - u^*\|_2^2 - \gamma^2\|w - w^*\|_2^2$$

for any u (either open or closed loop) and $w \in \mathcal{L}^2[0, T]$ in (2.1), noting that $x(0) = 0$. With $u = u^$, we have*

$$(2.10) \quad \|\mathcal{J}_{zw}\| < \gamma.$$

Proof. Since (2.6) is assumed to have a solution on $[0, T]$ with $P(T) = 0$, we have, for any u and w , and $x(0) = 0$,

$$\begin{aligned} J(u, w) &= \int_0^T \left\{ z'z - \gamma^2 w'w + \frac{d}{dt}(x'Px) \right\} dt \\ &= \int_0^T \{ x'C'Cx + u'u - \gamma^2 w'w + (x'A' + w'B_1' + u'B_2')Px \\ &\quad + x'\dot{P}x + x'P(Ax + B_1w + B_2u) \} dt. \end{aligned}$$

Substituting from (2.6) gives us

$$\begin{aligned} J(u, w) &= \int_0^T \left\{ x'P(B_2B_2' - \gamma^{-2}B_1B_1')Px + u'u - \gamma^2 w'w \right. \\ &\quad \left. + [w' \quad u'] \begin{bmatrix} B_1' \\ B_2' \end{bmatrix} Px + x'P \begin{bmatrix} B_1 & B_2 \end{bmatrix} \begin{bmatrix} w \\ u \end{bmatrix} \right\} dt \\ (2.11) \quad &= \int_0^T [u' + x'PB_2][u + B_2'Px] dt \\ &\quad - \int_0^T [w' - x'\gamma^{-2}PB_1]\gamma^2[w - \gamma^{-2}B_1'Px] dt \\ &= \|u - u^*\|_2^2 - \gamma^2\|w - w^*\|_2^2 \end{aligned}$$

with w^*, u^* as in (2.7) and (2.8), which establishes (2.9). Suppose that \mathcal{L} is the operator with realization

$$\dot{x} = (A + B_2K)x + B_1w, \quad (w - w^*) = -\gamma^{-2}B_1'Px + w,$$

which maps w to $(w - w^*)$. Since \mathcal{L}^{-1} exists (and is given by $\hat{x} = (A + B_2K + \gamma^{-2}B_1B_1'P)x + B_1(w - w^*)$, $w = \gamma^{-2}B_1'Px + (w - w^*)$), we can write

$$\|\mathcal{T}_{zw}w\|_2^2 - \gamma^2\|w\|_2^2 = -\gamma^2\|w - w^*\|_2^2 = -\gamma^2\|\mathcal{L}w\|_2^2 \leq -\kappa\|w\|_2^2$$

for some positive κ . Thus $\|\mathcal{T}_{zw}\| < \gamma$, as required. \square

Remark 2.1 (Positive semidefiniteness of $P(t)$). It is not hard to show that if (2.6) with $P(T) = 0$ has a solution on $[0, T]$, then $P(t) \geq 0$ for all $t \in [0, T]$. Consider system (2.1) with $0 \leq t_o \leq t \leq T$, and with $x(t_o) = x_o$. If

$$(2.12) \quad J_{t_o}(u, w) := \int_{t_o}^T (z'z - \gamma^2 w'w) dt,$$

then we may complete the square as above to yield

$$J_{t_o}(u, w) - x_o'P(t_o)x_o = \|u - u^*\|_2^2 - \gamma^2\|w - w^*\|_2^2,$$

in which x_o is regarded as an arbitrary initial condition for the running period $[t_o, T] \subseteq [0, T]$. Consequently,

$$(2.13) \quad J_{t_o}(u^*, 0) + \gamma^2\|w^*\|_2^2 = x_o'P(t_o)x_o$$

for every x_o . Since $J_{t_o}(u^*, 0) \geq 0$ for any x_o , it follows that $P(t) \geq 0$ for all $t \in [0, T]$.

Note that the game Riccati equation (2.6) and the corresponding linear quadratic regulator Riccati equation “approach each other” as γ increases. Since the linear quadratic regulator Riccati equation always has a solution, we expect (2.6) to have a solution if γ is large enough. In the next section, we show that the existence of an admissible controller satisfying (2.5) is a necessary condition for (2.6) to have a solution on $[0, T]$.

2.1. The necessary conditions. The aim of this section is to show that if an admissible feedback control exists that solves the \mathcal{H}^∞ -control problem, then (2.6) has a solution on $[0, T]$. Our proof uses conjugate point arguments that are reminiscent of those found in [25], [26]. To begin, we introduce the two-point boundary value problem

$$(2.14) \quad \begin{bmatrix} \dot{x} \\ \dot{p} \end{bmatrix} = \begin{bmatrix} A & -(B_2B_2' - \gamma^{-2}B_1B_1') \\ -C'C & -A' \end{bmatrix} \begin{bmatrix} x \\ p \end{bmatrix}, \quad \begin{bmatrix} x(0) \\ p(T) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

and its associated transition matrix, which is generated by the linear differential equation

$$(2.15) \quad \begin{bmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{bmatrix}(t, T) = \begin{bmatrix} A & -(B_2B_2' - \gamma^{-2}B_1B_1') \\ -C'C & -A' \end{bmatrix} \begin{bmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{bmatrix}(t, T),$$

with $\Phi(T, T) = I$. Next, we note that

$$\begin{aligned} -\frac{d}{dt}(\Phi_{21}\Phi_{11}^{-1}) &= -\dot{\Phi}_{21}\Phi_{11}^{-1} + \Phi_{21}\Phi_{11}^{-1}\dot{\Phi}_{11}\Phi_{11}^{-1} \\ &= A'(\Phi_{21}\Phi_{11}^{-1}) + (\Phi_{21}\Phi_{11}^{-1})A - (\Phi_{21}\Phi_{11}^{-1})(B_2B_2' - \gamma^{-2}B_1B_1')(\Phi_{21}\Phi_{11}^{-1}) \\ &\quad + C'C. \end{aligned}$$

Thus, if $\Phi_{11}^{-1}(t, T)$ exists on $t \in [0, T]$, the RDE (2.6) has a solution on $[0, T]$ given by $P(t) = \Phi_{21}(t, T)\Phi_{11}^{-1}(t, T)$.

The existence of $\Phi_{11}^{-1}(t, T)$ is equivalent to a conjugate point condition, and this observation forms the basis of the necessity proof.

DEFINITION 2.1. Suppose that $t_o < t_f$. Then t_o and t_f are *conjugate points* if it is possible to find a *nontrivial* solution to (2.14) such that $x(t_o) = p(t_f) = 0$. We note also that $p(t_o) \neq 0$ and $x(t_f) \neq 0$, since otherwise (2.14) would only have the trivial solution $[x' \ p'](t) \equiv 0$.

LEMMA 2.2. The matrix $\Phi_{11}(t_o, t_f)$ defined by (2.15) is singular if and only if t_o and t_f are conjugate points.

Proof. Suppose that t_o and t_f are conjugate points. Since

$$\begin{bmatrix} x(t_o) \\ p(t_o) \end{bmatrix} = \begin{bmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{bmatrix} (t_o, t_f) \begin{bmatrix} x(t_f) \\ p(t_f) \end{bmatrix},$$

and since t_o and t_f are conjugate points, we obtain

$$(2.16) \quad 0 = \Phi_{11}(t_o, t_f)x(t_f).$$

Hence $\Phi_{11}(t_o, t_f)$ is singular, as $x(t_f)$ is nonzero.

If, on the other hand, $\Phi_{11}(t_o, t_f)$ is singular, there exists a vector $g \neq 0$ such that $\Phi_{11}(t_o, t_f)g = 0$. By considering the final value problem with $x(t_f) = g$ and $p(t_f) = 0$, we see that $x(t_o) = 0$, which establishes that t_o and t_f are conjugate. \square

THEOREM 2.3. Consider the linear system (2.1). If there exists a closed loop control $\bar{K} \in \mathcal{C}$ such that

$$(2.17) \quad \|\mathcal{T}_{zw}\| < \gamma,$$

then $\Phi_{11}(t, T)$ is nonsingular for all $t \in [0, T]$. Consequently, the RDE (2.6), with boundary condition $P(T) = 0$, has a solution on $[0, T]$.

Proof. We suppose for contradiction that $t^* \in [0, T]$ is the largest time such that $\Phi_{11}(t^*, T)$ is singular. Since $\Phi_{11}(T, T) = I$, $t^* < T$ by continuity. It follows from Lemma 2.2 that t^* and T are conjugate points, giving $x(t^*) = p(T) = 0$. Next, we suppose that

$$\bar{w}(t) = \begin{cases} \gamma^{-2} B_1' p(t), & t^* \leq t \leq T, \\ 0, & 0 \leq t < t^*. \end{cases}$$

Since (2.14) has a nontrivial solution, $p(t) \neq 0$ $t \in [t^*, T]$. Furthermore, if $B_1' p(t) \equiv 0$, then (under this assumption) (2.14) becomes the two-point boundary problem for the LQ regulator; this can have no conjugate points [1]. Thus $\bar{w}(t) \neq 0$. Next, we see that $J(\bar{K}, \bar{w}) \geq J_{t^*}(\bar{K}, \bar{w})$, since $\bar{w}(t) = 0$ for $t < t^*$. Let $\tilde{u}(t)$ be the function generated by $x(t^*) = 0$, \bar{K} and $\bar{w}(t)$. Then it follows that

$$(2.18a) \quad J_{t^*}(\bar{K}, \bar{w}) = J_{t^*}(\tilde{u}, \bar{w}) \geq \min_{u(t)} \int_{t^*}^T \{ \tilde{x}' C' C \tilde{x} + u' u - \gamma^2 \bar{w}' \bar{w} \} dt,$$

subject to

$$(2.18b) \quad \dot{\tilde{x}} = A \tilde{x} + B_1 \bar{w} + B_2 u, \quad \tilde{x}(t^*) = 0.$$

The tilde is used to distinguish the state trajectory associated with (2.14) from that associated with the minimization problem in (2.18). The initial condition $\tilde{x}(t^*) = 0$ is a consequence of (i) $x(0) = 0$, (ii) $\bar{w} \equiv 0$ for all $t \in [0, t^*)$, and (iii) $K \in \mathcal{C}$. The minimization on the right-hand side of (2.18a) subject to (2.18b), with $u(\cdot)$ being sought in *open-loop* form is almost a standard LQ control problem and is solved as follows: Form

$$H(t, \tilde{x}, \lambda, u) = \frac{1}{2} (\tilde{x}' C' C \tilde{x} + u' u - \gamma^2 \bar{w}' \bar{w}) + \lambda (A \tilde{x} + B_1 \bar{w} + B_2 u).$$

Then

$$\frac{\partial H}{\partial u} = 0 \Rightarrow u_{\min} = -B_2' \lambda$$

and

$$-\dot{\lambda} = \frac{\partial H}{\partial x} \Rightarrow -\dot{\lambda} = C' C \tilde{x} + A' \lambda, \quad \lambda(T) = 0,$$

giving

$$(2.19) \quad \begin{bmatrix} \dot{\tilde{x}} \\ \dot{\lambda} \end{bmatrix} = \begin{bmatrix} A & -B_2 B_2' \\ -C' C & -A' \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \lambda \end{bmatrix} + \begin{bmatrix} B_1 \bar{w} \\ 0 \end{bmatrix}, \quad \begin{bmatrix} \tilde{x}(t^*) \\ \lambda(T) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Using $\bar{w} = \gamma^{-2} B_1' p$, and subtracting (2.19) from (2.14), gives us

$$(2.20) \quad \begin{bmatrix} \dot{\tilde{x}} - \dot{\hat{x}} \\ \dot{\lambda} - \dot{\hat{p}} \end{bmatrix} = \begin{bmatrix} A & -B_2 B_2' \\ -C' C & -A' \end{bmatrix} \begin{bmatrix} \tilde{x} - \hat{x} \\ \lambda - \hat{p} \end{bmatrix}, \quad \begin{bmatrix} (\tilde{x} - \hat{x})(t^*) \\ (\lambda - \hat{p})(T) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Since there are no conjugate points associated with the LQ boundary problem in (2.20), $\tilde{x}(t) \equiv x(t)$ and $\lambda(t) \equiv p(t)$ for all $t \in [t^*, T]$. Consequently,

$$\begin{aligned} \min_{u(t)} \int_{t^*}^T \{ \tilde{x}' C' C \tilde{x} + u' u - \gamma^2 \bar{w}' \bar{w} \} dt &= \int_{t^*}^T \{ \tilde{x}' C' C \tilde{x} + \lambda' B_2 B_2' \lambda - \gamma^{-2} p' B_1 B_1' p \} dt \\ &= \int_{t^*}^T \{ x' C' C x + p' B_2 B_2' p - \gamma^{-2} p' B_1 B_1' p \} dt \\ &= - \int_{t^*}^T \{ x' (\dot{\hat{p}} + A' p) + p' (\dot{\hat{x}} - A x) \} dt \text{ by (2.14)} \\ &= - \int_{t^*}^T \frac{d}{dt} \{ x' p \} dt = 0, \end{aligned}$$

which contradicts (2.17). \square

Remark 2.2. In our later work, we will need to consider problems in which the output is given by

$$(2.21) \quad z(t) = C(t)x(t) + D(t)u(t),$$

rather than (2.2). This change in output has the effect of introducing a cross-coupling term between u and x in the functional $J(K, w)$. This may be removed by the change of variable $u = \hat{u} - D' C x$. In the case of an output given by (2.21), the Riccati equation (2.6) becomes

$$\begin{aligned} -\dot{P} &= P(A - B_2 D' C) + (A - B_2 D' C)P - P(B_2 B_2' - \gamma^{-2} B_1 B_1')P \\ &\quad + C(I - DD')C, \quad P(T) = 0, \end{aligned}$$

and the corresponding minimizing feedback control is given by $u^*(t) = -(D' C + B_2' P)(t)x(t)$, rather than (2.7).

2.2. A representation formula for all solutions. In Theorem 2.1 we found one feedback control that solves the time-varying \mathcal{H}^∞ -control problem. In the infinite-horizon case, it is well known that there are usually many feedback controls that result in $\|\mathcal{T}_{zw}\| < \gamma$. The aim of this section is to show how to construct all linear full-information (access to w and x) feedback controls such that $\|\mathcal{T}_{zw}\| < \gamma$. Our analysis is based on the output equation (2.21), rather than (2.2).

THEOREM 2.4. *Suppose that*

$$(2.22a) \quad \dot{\hat{x}}(t) = A(t)x(t) + B_1(t)w(t) + B_2(t)u(t), \quad x(0) = 0,$$

$$(2.22b) \quad z(t) = C(t)x(t) + D(t)u(t)$$

is given. Then there exists a control $\bar{K} \in \mathcal{C}$ such that $\|\mathcal{T}_{zw}\| < \gamma$ if and only if

$$(2.23) \quad -\dot{P} = (A - B_2 D' C)' P + P(A - B_2 D' C) - P(B_2 B_2' - \gamma^{-2} B_1 B_1') P + C'(I - DD')C$$

has a solution on $[0, T]$ with $P(T) = 0$. In this case, all closed-loop systems \mathcal{T}_{zw} generated by controls of the form

$$(2.24) \quad u(t) = -(\mathfrak{L}_1 x + \mathfrak{L}_2 w)(t),$$

where \mathfrak{L}_1 and \mathfrak{L}_2 are arbitrary causal linear operators such that

$$\|\mathcal{T}_{zw}\| < \gamma$$

are also generated by

$$(2.26) \quad u(t) = u^*(t) + (\mathfrak{U}(w - w^*))(t),$$

where

$$(2.27a) \quad u^*(t) = -F_\infty x = -(D' C + B_2' P)x,$$

$$(2.27b) \quad w^*(t) = \gamma^{-2} B_1' P x,$$

and $\gamma^{-1} \mathfrak{U}$ is an arbitrary linear causal strictly contractive operator on $\mathcal{L}^2[0, T]$ (i.e., $\|\mathfrak{U}\| < \gamma$).

Equivalently, every \mathfrak{L}_1 and \mathfrak{L}_2 may be parametrized by

$$(2.28a) \quad \mathfrak{L}_1 = F_\infty + \gamma^{-2} \mathfrak{U} B_1' P,$$

$$(2.28b) \quad \mathfrak{L}_2 = -\mathfrak{U}.$$

Equations (2.22), (2.26), and (2.27) may be represented diagrammatically, as in Fig. 1.

Proof. The first part, that the existence of a controller such that $\|\mathcal{T}_{zw}\| < \gamma$ is necessary and sufficient for the existence of $P(t)$ is just Theorems 2.1 and 2.3 for the cross-coupled game.

Now suppose that the RDE (2.23) with $P(T) = 0$ has a solution on $[0, T]$. We need to show (1) that there exists a causal \mathfrak{U} in (2.26) corresponding to every control of the form given in (2.24), and (2) that $\|\mathcal{T}_{zw}\| < \gamma$ if and only if $\gamma^{-1} \mathfrak{U}$ is strictly contractive.

Substituting (2.24) and (2.27b) into the dynamical equation (2.22a) gives us

$$\dot{x} = (A + \gamma^{-2} B_1 B_1' P - B_2(\mathfrak{L}_1 + \gamma^{-2} \mathfrak{L}_2 B_1' P))x + (B_1 - B_2 \mathfrak{L}_2)(w - w^*), \quad x(0) = 0,$$

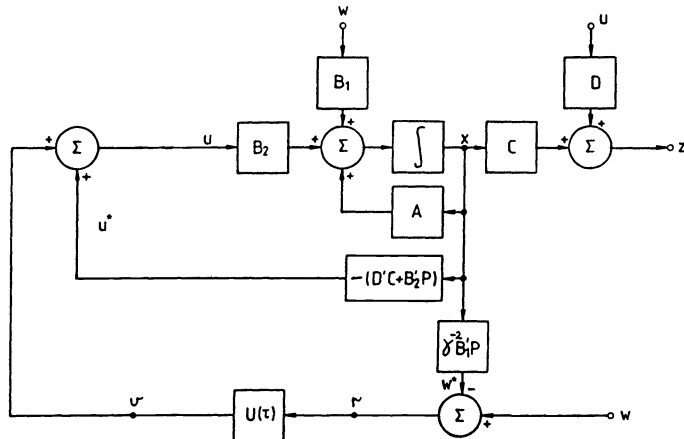


FIG. 1. All solutions to the \mathcal{H}^∞ -control problem with perfect information.

which shows that x depends causally on $(w - w^*)$, and so $x(t) = \mathfrak{L}_3(w - w^*)$, in which \mathfrak{L}_3 is a linear causal operator. Substituting (2.27) into (2.24) gives us

$$\begin{aligned} u - u^* &= -(\mathfrak{L}_1 - F_\infty - \gamma^{-2}\mathfrak{L}_2 B_1' P)x - \mathfrak{L}_2(w - w^*) \\ &= [-(\mathfrak{L}_1 - F_\infty - \gamma^{-2}\mathfrak{L}_2 B_1' P)\mathfrak{L}_3 - \mathfrak{L}_2](w - w^*) \\ &= \mathfrak{U}(w - w^*) \end{aligned}$$

for some linear causal \mathfrak{U} . This establishes the existence of the causal mapping in (2.26). The proof is completed by noting that $\gamma^{-1}\mathfrak{U}$ is strictly contractive as follows:

$$\begin{aligned} \|z\|_2^2 - \gamma^2\|w\|_2^2 &= \|u - u^*\|_2^2 - \gamma^2\|w - w^*\|_2^2 \\ &= \|\mathfrak{U}(w - w^*)\|_2^2 - \gamma^2\|w - w^*\|_2^2 \\ &= \|\mathfrak{U}\mathcal{L}w\|_2^2 - \gamma^2\|\mathcal{L}w\|_2^2 \\ &< 0 \text{ for all } w \neq 0 \in \mathcal{L}^2[0, T] \Leftrightarrow \gamma^{-1}\mathfrak{U} \text{ is strictly contractive.} \end{aligned}$$

In the above, \mathcal{L} is the causal and causally invertible operator linking w and $(w - w^*)$. The invertibility of \mathcal{L} was established in the proof of Theorem 2.1. \square

At this point it is interesting to reexamine Fig. 1. By reviewing (2.26), we see that $r = (w - w^*)$, and that $u = (v + u^*)$ drives B_2 with $v = \mathfrak{U}r$. If $w = w^*$, there is no signal into the \mathfrak{U} parameter, and the corresponding control is given by u^* (irrespective of \mathfrak{U}). If $w \neq w^*$, we do not have to use the control u^* . Thus for the purpose of solving the \mathcal{H}^∞ -control problem, the u -player only has to play well enough to ensure that $J(K, w) < 0$ for a given $w \neq 0 \in \mathcal{L}^2[0, T]$.

2.3. Adjoint systems. In § 3, where we derive a representation formula for all solutions to the time-varying \mathcal{H}^∞ problem with output feedback, we will require an elementary property of adjoint systems [8], [14].

Let $\mathfrak{L}: X \rightarrow Y$ be a linear operator between two Hilbert spaces X and Y . Then $\mathfrak{L}^*: Y \rightarrow X$, the adjoint of \mathfrak{L} , is the unique linear operator such that, for all $z \in Y$ and $w \in X$, $\langle z, \mathfrak{L}w \rangle = \langle \mathfrak{L}^*z, w \rangle$ [14, Thm. 2, p. 39]. Note also that $\|\mathfrak{L}\| = \|\mathfrak{L}^*\|$. Now suppose that \mathfrak{L} is described by the state-space equations

$$(2.29a) \quad \dot{x}(t) = A(t)x(t) + B(t)w(t), \quad x(0) = 0,$$

$$(2.29b) \quad y(t) = C(t)x(t) + D(t)w(t),$$

or, equivalently,

$$(2.30) \quad \begin{bmatrix} \frac{d}{dt} I - A(t) & -B(t) \\ C(t) & D(t) \end{bmatrix} \begin{bmatrix} x(t) \\ w(t) \end{bmatrix} = \begin{bmatrix} 0 \\ y(t) \end{bmatrix}, \quad x(0) = 0.$$

Thus

$$\begin{aligned} \langle z, \mathfrak{L}w \rangle &= \left\langle \begin{bmatrix} p \\ z \end{bmatrix}, \begin{bmatrix} \frac{d}{dt} I - A & -B \\ C & D \end{bmatrix} \begin{bmatrix} x \\ w \end{bmatrix} \right\rangle \\ &= \int_0^T \left(p' \frac{dx}{dt} \right) dt - \int_0^T p'(Ax + Bw) dt + \int_0^T z'(Cx + Dw) dt. \end{aligned}$$

Integrating the first term by parts gives us

$$\begin{aligned}
 &= p'(t_1)x(t_1) - \int_0^T \left(\frac{dp'}{dt} \right) x dt - \int_0^T p'(Ax + Bw) dt \\
 &\quad + \int_0^T z'(Cx + Dw) dt \\
 &= \left\langle \begin{bmatrix} -\left(\frac{d}{dt} I + A' \right) & C' \\ -B' & D' \end{bmatrix} \begin{bmatrix} p \\ z \end{bmatrix}, \begin{bmatrix} x \\ w \end{bmatrix} \right\rangle + p'(T)x(T) \\
 &= \langle \mathfrak{L}^* z, w \rangle,
 \end{aligned}$$

where the adjoint \mathfrak{L}^* is a linear system with realization

$$(2.31a) \quad -\dot{p}(t) = A'(t)p(t) + C'(t)z(t), \quad p(T) = 0,$$

$$(2.31b) \quad q(t) = B'(t)p(t) + D'(t)z(t).$$

In the following, we will require the solution of a dual LQ game system, which is obtained by applying Theorem 2.4 to its associated adjoint system.

2.4. The asymptotic properties of the game equation. This section represents a digression from the main stream of the time-varying finite-horizon \mathcal{H}^∞ -control problem, and its purpose is to make some connections with the more familiar time-invariant, infinite interval results [9]. In particular, we show that in the limit as $T \rightarrow \infty$, the solution of (2.6) approaches the smallest nonnegative solution of an algebraic Riccati equation (if such a solution exists). We also show that this smallest nonnegative solution is stabilizing in the sense alluded to in [9]. We begin by connecting the properties of the solution of the RDE (2.6) as $T \rightarrow \infty$ (when the various coefficient matrices are assumed to be constant) to the solution of the algebraic Riccati equation

$$(2.32) \quad 0 = PA + A'P - P(B_2B_2' - \gamma^{-2}B_1B_1')P + C'C.$$

In doing this, we will suppose that each of the matrices in (2.1) and (2.2) is time-invariant and that (A, B_2, C) is stabilizable and observable (it is not hard to remove the observability assumption, but we will not address this issue here).

If (2.6) is solved backward from the terminal condition $P(T) = Q$, then we will refer to the solution as $P(t, T, Q)$. If a limiting solution to (2.6) exists, we will call it $\bar{P}(t) = \lim_{T \uparrow \infty} P(t, T, Q)$. The first result of this section is standard and shows that $P(t)$ is nonincreasing if $P(T) = 0$.

LEMMA 2.5. *If (2.6) has a solution on $[0, T]$ with $P(T) = 0$, then $P(t)$ is nonincreasing (in the sense of semidefinite matrices).*

Proof. Differentiating (2.6) gives us

$$-\dot{\bar{P}} = \dot{\bar{P}}(A - (B_2B_2' - \gamma^{-2}B_1B_1')P) + (A - (B_2B_2' - \gamma^{-2}B_1B_1')P)' \dot{\bar{P}}, \quad \dot{\bar{P}}(T) = -C'C,$$

which has solution [8, Thm. 1, p. 59] $\dot{\bar{P}}(t) = -\Phi(t, T)'C'C\Phi(t, T)$, where $\Phi(t, T)$ is the transition matrix associated with $(A - (B_2B_2' - \gamma^{-2}B_1B_1')P)'$. It is evident that $\dot{\bar{P}}(t) \leq 0$ for all t , so $P(t)$ is nonincreasing. \square

If (2.32) has at least one nonnegative solution, then we will show that the smallest of these solutions, Y say, is an upper bound for $P(t)$ provided that $Y \geq P(T)$.

LEMMA 2.6. *Suppose that $P(t, T, Q)$ exists $[0, T]$ and that $Y = Y' \geq 0$ satisfies (2.32) with $Y \geq Q$. Then $Y \geq P(t)$ for all $t \in [0, T]$.*

Proof. Completing the square using (2.32) gives us

$$J_t(u, w) = \int_t^T [(u + B_2' Yx)'(u + B_2' Yx) - \gamma^2(w - \gamma^{-2} B_1' Yx)'(w - \gamma^{-2} B_1' Yx)] d\tau \\ + x'(t) Yx(t) - x'(T) Yx(T)$$

for any u and w . In the same way,

$$J_t(u, w) = \int_t^T [(u + B_2' Px)'(u + B_2' Px) - \gamma^2(w - \gamma^{-2} B_1' Px)'(w - \gamma^{-2} B_1' Px)] d\tau \\ + x'(t) P(t)x(t) - x'(T) P(T)x(T)$$

by invoking (2.6). Subtracting and setting $\tilde{u} := -B_2' Yx$ and $\tilde{w} := \gamma^{-2} B_1' Px$ yields

$$x'(t)[Y - P(t)]x(t) = x'(T)[Y - P(T)]x(T) + \gamma^2 \int_t^T (\tilde{w} - \gamma^{-2} B_1' Yx)'(\tilde{w} - \gamma^{-2} B_1' Yx) d\tau \\ + \int_t^T (\tilde{u} + B_2' Px)'(\tilde{u} + B_2' Px) d\tau \geq 0$$

for all $x(t)$. \square

We remark that a variation on the above argument, in conjunction with Lemma 2.5, will establish the following result.

COROLLARY 2.7. *Assume that $Y = Y' \geq 0$ satisfies (2.32) with $Y \geq Q \geq 0$. Then $P(t, T, Q)$ exists for all $t \leq T$ and $Y \geq P(t, T, Q) \geq P(t, T, 0) \geq 0$.*

If $\bar{P}(t)$ exists for some Q , it is easy to argue that it satisfies (2.6). Suppose that $t \leq T_1 \leq T$. Then $P(t, T, Q) = P(t, T_1, P(T_1, T, Q))$, and thus

$$\bar{P}(t) = \lim_{T \uparrow \infty} P(t, T, Q) = \lim_{T \uparrow \infty} P(t, T_1, P(T_1, T, Q)).$$

For any fixed T_1 , the solution $P(t, T_1, \tilde{Q})$ depends continuously on \tilde{Q} and therefore

$$(2.33) \quad \bar{P}(t) = P(t, T_1, \lim_{T \uparrow \infty} P(T_1, T, Q)) = P(t, T_1, \bar{P}(T_1)),$$

which shows that $\bar{P}(t)$ is a solution to (2.6) for all t . Since the system and output matrices have been assumed to be time-invariant, the value of the game $J_t(K^*, w^*) = x'(t)P(t)x(t)$ is invariant under time translations. Consequently, $\bar{P}(t) = \bar{P}$ is constant and therefore satisfies the algebraic equation (2.32).

If a nonnegative stabilizing (i.e., $(A - (B_2 B_2' - \gamma^{-2} B_1 B_1')P)$ asymptotically stable) solution to (2.32) exists, it is the smallest of the nonnegative solutions.

THEOREM 2.8. *Suppose that $[A, C]$ is observable and that $X \geq 0$ and $Y \geq 0$ satisfy*

$$(2.34) \quad A'P + PA - PDP + C'C = 0,$$

with $A - DX$ asymptotically stable. Then $Y \geq X$.

Proof. Due to the observability of $[A, C]$, every solution to (2.32) is nonsingular. We can therefore write $X(A - DX)X^{-1} = -(A' + C'CX^{-1})$, which shows that $A' + C'CX^{-1}$ is completely unstable. Now

$$AX^{-1} + X^{-1}A' - D + X^{-1}C'CX^{-1} = 0 \quad \text{and} \quad AY^{-1} + Y^{-1}A' - D + Y^{-1}C'CY^{-1} = 0.$$

Subtracting gives us

$$A(X^{-1} - Y^{-1}) + (X^{-1} - Y^{-1})A' + X^{-1}C'CX^{-1} - Y^{-1}C'CY^{-1} = 0 \\ \Rightarrow (A + X^{-1}C'C)(X^{-1} - Y^{-1}) + (X^{-1}Y^{-1})(A + X^{-1}C'C)' \\ = (X^{-1} - Y^{-1})C'C(X^{-1} - Y^{-1}).$$

Since $(A + X^{-1}C'C)$ is completely unstable, we must have $(X^{-1} - Y^{-1}) \geq 0$ and so $X \leq Y$. \square

Remark 2.3. Theorem 2.8 remains valid when the observability assumption on $[A, C]$ is weakened to one of detectability. Suppose that

$$A = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix}, \quad C = [C_1 \ 0],$$

with $[A_{11}, C_1]$ observable, then argue as before on the smaller system $[A_{11}, C_1]$.

3. Solution to the time-varying \mathcal{H}^∞ -control problem: Main results. Now that the background game theory is in place, we are in a position to solve the general time-varying \mathcal{H}^∞ -control problem. As mentioned in the Introduction, the plant is described by

$$(3.1a) \quad \dot{x}(t) = A(t)x(t) + B_1(t)w(t) + B_2(t)u(t), \quad x(0) = 0,$$

$$(3.1b) \quad z(t) = C_1(t)x(t) + D_{12}(t)u(t),$$

$$(3.1c) \quad y(t) = C_2(t)x(t) + D_{21}(t)w(t),$$

in which we assume that all the matrices have entries that are continuous functions of time, and that D_{12} and D_{21} have full column and row rank, respectively, for all $t \in [0, T]$. Under this assumption, we may scale the problem so that $D'_{12}D_{12} = I$ and $D_{21}D'_{21} = I$. By a corollary of Doležal's theorem [29, Cor. 3, p. 70], we know that there exist continuous extensions D_\perp and \tilde{D}_\perp to D_{12} and D_{21} , respectively, such that $[D_{12} \ D_\perp]$ and $[D_{21}^* \ \tilde{D}_\perp^*]$ are orthogonal for all $t \in [0, T]$. Note that by generalizing the loop-shifting transformations in [11], [24] to the time-varying case, there is no loss of generality in the assumption that D_{11} and D_{22} are zero. Details of the scaling and loop-shifting transformations required in the time-varying case may be found in the Appendix.

3.1. Problems of the first kind. We begin with a preliminary result in which we assume that D_{12} and D_{21} are square; we call such problems problems of the first kind. Under this assumption, scaling arguments allow us to assume, without loss of generality, that $D_{12} = I$ and $D_{21} = I$. As we will now show, finite-horizon \mathcal{H}^∞ problems of the first kind may be solved by inverting the plant, and the resulting controllers have a remarkably simple observer structure.

THEOREM 3.1. *Consider the generalized plant described by*

$$(3.2a) \quad \dot{x}(t) = A(t)x(t) + B_1(t)w(t) + B_2(t)u(t), \quad x(0) = 0,$$

$$(3.2b) \quad z(t) = C_1(t)x(t) + u(t),$$

$$(3.2c) \quad y(t) = C_2(t)x(t) + w(t).$$

Then (i) the set of all linear causal output-feedback control laws such that $\|\mathcal{F}_{zw}\| < \gamma$ is parametrized by

$$(3.3a) \quad \dot{\hat{x}}(t) = (A - B_1C_2 - B_2C_1)(t)\hat{x}(t) + B_1(t)y(t) + B_2(t)v(t), \quad \hat{x}(0) = 0,$$

$$(3.3b) \quad u(t) = v(t) - C_1(t)\hat{x}(t),$$

$$(3.3c) \quad r(t) = y(t) - C_2(t)\hat{x}(t),$$

$$(3.3d) \quad v(t) = (\mathbb{U}r)(t),$$

in which \mathbb{U} is a causal strictly contractive operator on $\mathcal{L}^2[0, T]$, and (ii) there exists a linear causal controller such that the closed-loop mapping from w to z is identically zero.

Proof. Eliminating $y(t)$ and $u(t)$ from (3.2) and (3.3) gives us

$$\begin{bmatrix} \dot{\hat{x}} \\ \dot{\hat{x}} \end{bmatrix} = \begin{bmatrix} A - B_1 C_2 - B_2 C_1 & B_1 C_2 \\ -B_2 C_1 & A \end{bmatrix} \begin{bmatrix} \hat{x} \\ x \end{bmatrix} + \begin{bmatrix} B_1 & B_2 \\ B_1 & B_2 \end{bmatrix} \begin{bmatrix} w \\ v \end{bmatrix}$$

and

$$\begin{bmatrix} z \\ r \end{bmatrix} = \begin{bmatrix} -C_1 & C_1 \\ -C_2 & C_2 \end{bmatrix} \begin{bmatrix} \hat{x} \\ x \end{bmatrix} + \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} \begin{bmatrix} w \\ v \end{bmatrix}.$$

Consequently,

$$(3.4) \quad \begin{bmatrix} \dot{\hat{x}} \\ \dot{\hat{x}} - \dot{\hat{x}} \end{bmatrix} = \begin{bmatrix} A - B_2 C_1 & B_1 C_2 \\ 0 & A - B_1 C_2 \end{bmatrix} \begin{bmatrix} \hat{x} \\ x - \hat{x} \end{bmatrix} + \begin{bmatrix} B_1 & B_2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} w \\ v \end{bmatrix}$$

and

$$\begin{bmatrix} z \\ r \end{bmatrix} = \begin{bmatrix} 0 & C_1 \\ 0 & C_2 \end{bmatrix} \begin{bmatrix} \hat{x} \\ x - \hat{x} \end{bmatrix} + \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} \begin{bmatrix} w \\ v \end{bmatrix}.$$

The second row of (3.4) implies that the observations error $e(t) = (x - \hat{x})(t)$ is identically zero, thereby establishing the observer property.¹ In addition, it is clear that $\begin{bmatrix} z \\ r \end{bmatrix}(t) = \begin{bmatrix} v \\ w \end{bmatrix}(t)$ for all $t \in [0, T]$, which shows that the controller is plant inverting.

For the first part, we note that if \mathcal{R} is any strictly contractive mapping from $w(t)$ to $z(t)$, then \mathcal{R} will be generated by setting $\mathfrak{U} = \mathcal{R}$ in (3.3d). Setting $\mathfrak{U} = 0$ proves the second part. \square

The controller given in Theorem 3.1 is presented diagrammatically in Fig. 2, and is seen to have an observer structure with observer gain matrix $-B_1$ and state estimate feedback C_1 .

3.2. Problems of the second kind. In this section we show that all problems in which either D_{12} or D_{21} is square, which we will refer to as problems of the second kind, have controllers that may be characterized by the solution of a single RDE. Simple plant inversion is no longer possible, since this requires that both D_{12} and D_{21} be square. Nevertheless, a solution is still possible and involves the game theory of § 2. If D_{21} is assumed to be square, we can replace (3.1c) with

$$(3.1d) \quad y(t) = C_2(t)x(t) + w(t)$$

by an appropriate scaling operation.

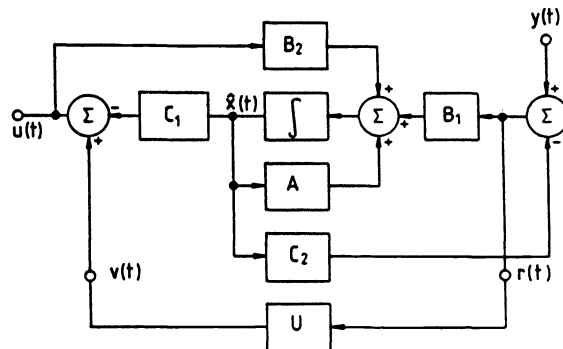


FIG. 2. The plant inverse as an observer.

¹ Since we are dealing with a finite time problem, the question of stability does not arise.

We now consider the possibility of obtaining a controller via a combination of an observer for x , and a linear feedback law as computed in § 2. Since the controllers that we are considering do not have access to w , the observer can only be driven by u and y . Suppose that

$$(3.5) \quad \dot{\hat{x}} = A\hat{x} + B_2u + K(C_2\hat{x} - y), \quad \hat{x}(0) = 0.$$

Subtracting (3.1) from (3.5) and substituting (3.1d) gives us

$$(3.6) \quad (\dot{\hat{x}} - \dot{x}) = A(\hat{x} - x) + K(C_2\hat{x} - C_2x - w) - B_1w = (A + KC_2)(\hat{x} - x) - (K + B_1)w.$$

Setting $K = -B_1$ as before, and $e := \hat{x} - x$ gives us

$$(3.7) \quad \dot{e} = (A - B_1C_2)e, \quad e(0) = 0$$

$\Rightarrow e(t) \equiv 0$ and so $\hat{x}(t) \equiv x(t)$. Again, the stability of $A - B_1C_2$ is not an issue for finite terminal times.

With the required observer property established, we may now find all linear closed-loop controls $K \in \mathcal{C}$ such that $\|\mathcal{T}_{zw}\| < \gamma$ as if full state information were available. As before,

$$(3.8) \quad J(u, w) = \int_0^T (z'z - \gamma^2 w'w) dt.$$

Using the results of § 2, we introduce the Riccati equation

$$(3.9) \quad \begin{aligned} -\dot{X}_\infty &= (A - B_2D'_{12}C_1)'X_\infty + X_\infty(A - B_2D'_{12}C_1) - X_\infty(B_2B'_2 - \gamma^{-2}B_1B'_1)X_\infty \\ &\quad + C'_1D_\perp D'_\perp C_1 \end{aligned}$$

with terminal condition $X_\infty(T) = 0$, and define

$$(3.10) \quad u^* = -(D'_{12}C_1 + B'_2X_\infty)x = -F_\infty x,$$

$$(3.11) \quad w^* = \gamma^{-2}B'_1X_\infty x,$$

which gives us

$$(3.12) \quad J(-F_\infty x, w) \leq -\varepsilon \|w\|_2^2$$

for some $\varepsilon > 0$.

By combining the observer given in (3.5) with $K = -B_1$, the equilibrium strategies (3.10), (3.11), and the characterization of all solutions in § 2.2, we obtain the controller configuration illustrated in Fig. 3 and the main result of this section.

THEOREM 3.2. *Given*

$$(3.13a) \quad \dot{x}(t) = A(t)x(t) + B_1(t)w(t) + B_2(t)u(t), \quad x(0) = 0,$$

$$(3.13b) \quad z(t) = C_1(t)x(t) + D_{12}(t)u(t),$$

$$(3.13c) \quad y(t) = C_2(t)x(t) + w(t),$$

then (i) an output-feedback controller exists such that $\|\mathcal{T}_{zw}\| < \gamma$ if and only if the RDE (3.9) has a solution on $[0, T]$ with terminal condition $X_\infty(T) = 0$, and (ii) every closed-loop operator \mathcal{T}_{zw} with $\|\mathcal{T}_{zw}\| < \gamma$, corresponding to an output feedback control $u = \mathcal{R}y$, is generated by

$$(3.14a) \quad \dot{\hat{x}}(t) = (A - B_1C_2 - B_2F_\infty)(t)\hat{x}(t) + B_1(t)y(t) + B_2(t)v(t), \quad \hat{x}(0) = 0,$$

$$(3.14b) \quad u(t) = v(t) - F_\infty(t)\hat{x}(t),$$

$$(3.14c) \quad r(t) = y(t) - (C_2 + \gamma^{-2}B'_1X_\infty)(t)\hat{x}(t),$$

$$(3.14d) \quad v(t) = (\mathbb{U}r)(t),$$

in which $\gamma^{-1}\mathbb{U}$ is a causal, strictly contractive operator on $\mathcal{L}^2[0, T]$.

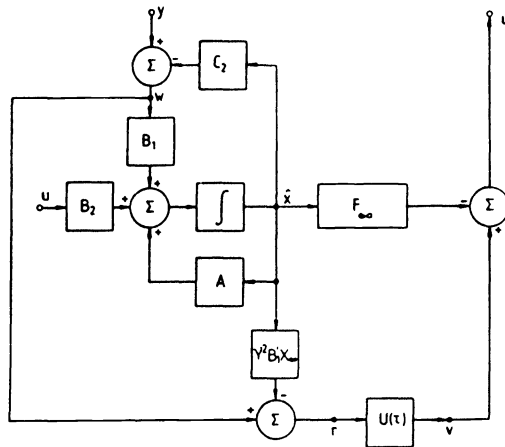


FIG. 3. All controllers for problems of the second kind.

Proof. (i) Suppose that a control of the form $u = \Re y$ exists such that $\|\mathcal{T}_{zw}\| < \gamma$. Then setting $u = \Re[C_2 \quad I][\begin{smallmatrix} x \\ w \end{smallmatrix}]$ in Theorem 2.3 proves the “only if” part. To prove the “if” part, we suppose that a solution to (3.9) exists. We then invoke the observer property developed in (3.5) to (3.7) and the completing-the-square argument given in Theorem 2.1.

Part (ii) is a consequence of Theorem 2.4 on substituting $\hat{x}(t)$ for $x(t)$. \square

A solution for problems of the second kind in which D_{12} , rather than D_{21} , is square is now given without proof. This particular result is needed in the next section in solutions to problems of the third kind and is solved via the dual game associated with the operator $[\mathfrak{P}_{11}^* \quad \mathfrak{P}_{21}^*]$ (see (1.1) for a definition).

THEOREM 3.3. *Given*

$$(3.15a) \quad \dot{x}(t) = A(t)x(t) + B_1(t)w(t) + B_2(t)u(t), \quad x(0) = 0,$$

$$(3.15b) \quad z(t) = C_1(t)x(t) + u(t),$$

$$(3.15c) \quad y(t) = C_2(t)x(t) + D_{21}w(t);$$

then (i) a feedback controller exists such that $\|\mathcal{T}_{zw}\| < \gamma$ if and only if the RDE

$$(3.16) \quad \begin{aligned} \dot{Y}_\infty = & (A - B_1 D_{21}' C_2) Y_\infty + Y_\infty (A - B_1 D_{21}' C_2)' - Y_\infty (C_2' C_2 - \gamma^{-2} C_1' C_1) Y_\infty \\ & + B_1 \tilde{D}_\perp' \tilde{D}_\perp B_1' \end{aligned}$$

has a solution on $[0, T]$ with $Y_\infty(0) = 0$, and (ii) every closed-loop operator \mathcal{T}_{zw} with $\|\mathcal{T}_{zw}\| < \gamma$, corresponding to an output feedback control $u = \Re y$, is generated by

$$(3.17a) \quad \dot{\hat{x}}(t) = (A - B_2 C_1 - H_\infty' C_2)(t) \hat{x}(t) - H_\infty'(t) y(t) - (B_2 + \gamma^{-2} Y_\infty C_1')(t) v(t),$$

$$(3.17b) \quad u(t) = C_1(t) \hat{x}(t) + v(t),$$

$$(3.17c) \quad r(t) = C_2(t) \hat{x}(t) + y(t),$$

$$(3.17d) \quad v(t) = (\mathbb{U}r)(t),$$

in which $\gamma^{-1} \mathbb{U}$ is a causal, strictly contractive operator on $\mathcal{L}^2[0, T]$ and $H_\infty = D_{21} B_1' + C_2 Y_\infty$.

Proof. Apply Theorem 3.2 to the adjoint system associated with (3.15). \square

3.3. Problems of the third kind. In this section we treat the case in which neither D_{21} nor D_{12} is square. We call such problems problems of the third kind. In the case where D_{21} has fewer rows than columns, the observer analysis in § 3.2 breaks down because the equation $\Re D_{21}(t) = -B_1(t)$ need not be soluble for \Re . What is required is a norm-preserving transformation to a new problem for which state reconstruction is possible. The desired transformation, which was first suggested in [9] in the time-invariant case, is immediate from identity (2.9). Suppose that

$$(3.18) \quad r = w - \gamma^{-2} B_1' X_\infty x,$$

$$(3.19) \quad v = u + F_\infty x.$$

Then (3.18), (3.19), and the completion-of-squares argument resulting in (2.9) gives us

$$(3.20) \quad \int_0^T (z'z - \gamma^2 w'w) dt = \int_0^T (v'v - \gamma^2 r'r) dt,$$

or, what is the same,

$$(3.21) \quad \|z\|_2^2 - \gamma^2 \|w\|_2^2 = \|v\|_2^2 - \gamma^2 \|r\|_2^2$$

for any r and v . Substituting (3.18) and (3.19) into (3.1) yields

$$(3.22a) \quad \dot{x} = (A + \gamma^{-2} B_1 B_1' X_\infty)x + B_1 r + B_2 u, \quad x(0) = 0,$$

$$(3.22b) \quad v = F_\infty x + Iu,$$

$$(3.22c) \quad y = (C_2 + \gamma^{-2} D_{21} B_1' X_\infty)x + D_{21} r.$$

LEMMA 3.4. *Suppose that the RDE (3.9) with $X_\infty(T) = 0$ has a solution on $[0, T]$. Suppose also that \Re is any controller, and that the control law $u = \Re y$ is applied to the systems given in (3.1) and (3.22)². Then $\|\mathcal{T}_{zw}\| < \gamma$ if and only if $\|\mathcal{T}_{vr}\| < \gamma$.*

Proof. First, note that the relationship between r and w in (3.18) is causally invertible. The result now follows directly from (3.21). \square

Lemma 3.4 shows that the tasks of designing controllers for the systems in (3.1) and (3.22) are interchangeable. The key feature of (3.22) is that D_{12} is square, and thus this system description is a problem of the second kind, which makes state reconstruction possible along the lines of § 3.2. If we define

$$(3.23) \quad F\left(\begin{bmatrix} \mathfrak{P}_{11} & \mathfrak{P}_{12} \\ \mathfrak{P}_{21} & \mathfrak{P}_{22} \end{bmatrix}, \Re\right) := \mathfrak{P}_{11} + \mathfrak{P}_{12} \Re (I - \mathfrak{P}_{22} \Re)^{-1} \mathfrak{P}_{21},$$

then it is immediate that

$$(3.24) \quad \left\{ F\left(\begin{bmatrix} \mathfrak{P}_{11} & \mathfrak{P}_{12} \\ \mathfrak{P}_{21} & \mathfrak{P}_{22} \end{bmatrix}, \Re\right) \right\}^* = F\left(\begin{bmatrix} \mathfrak{P}_{11}^* & \mathfrak{P}_{21}^* \\ \mathfrak{P}_{12}^* & \mathfrak{P}_{22}^* \end{bmatrix}, \Re^*\right)$$

and that

$$(3.25) \quad \left\| F\left(\begin{bmatrix} \mathfrak{P}_{11} & \mathfrak{P}_{12} \\ \mathfrak{P}_{21} & \mathfrak{P}_{22} \end{bmatrix}, \Re\right) \right\|_\infty = \left\| F\left(\begin{bmatrix} \mathfrak{P}_{11}^* & \mathfrak{P}_{21}^* \\ \mathfrak{P}_{12}^* & \mathfrak{P}_{22}^* \end{bmatrix}, \Re^*\right) \right\|_\infty.$$

A direct application of the results of § 2.3 (see (2.31)) shows that a state-space model for the adjoint operator associated with (3.22) is given by

$$(3.26a) \quad -\dot{p} = (A + \gamma^{-2} B_1 B_1' X_\infty)' p + F_\infty' u_1 + (C_2 + \gamma^{-2} D_{21} B_1' X_\infty)' u_2, \quad p(T) = 0,$$

$$(3.26b) \quad y_1 = B_1' p + D_{21}' u_2,$$

$$(3.26c) \quad y_2 = B_2' p + u_1.$$

² For the purposes of this result, \Re need not be linear.

If we now substitute (3.26) into (3.14), using (3.9) to (3.11), we obtain a representation formula for \mathfrak{R}^* and thus also for \mathfrak{R} . This calculation forms the basis of the next result.

THEOREM 3.5. *Given*

$$(3.27a) \quad \dot{x}(t) = A(t)x(t) + B_1(t)w(t) + B_2(t)u(t), \quad x(0) = 0,$$

$$(3.27b) \quad z(t) = C_1(t)x(t) + D_{12}(t)u(t),$$

$$(3.27c) \quad y(t) = C_2(t)x(t) + D_{21}(t)w(t),$$

then (i) a causal output-feedback control of the form $u = \mathfrak{R}y$ exists such that $\|\mathcal{T}_{zw}\| < \gamma$ if and only if the RDEs (3.9) and (3.30a) (given below) have solutions on $[0, T]$, and (ii) every closed-loop operator \mathcal{T}_{zw} with $\|\mathcal{T}_{zw}\| < \gamma$, corresponding to an output feedback control $u = \mathfrak{R}y$, is generated by

$$(3.28a) \quad \dot{x}_k(t) = A_k(t)x_k(t) + B_{k1}(t)y(t) + B_{k2}(t)v(t), \quad x_k(0) = 0,$$

$$(3.28b) \quad u(t) = C_{k1}(t)x_k(t) + v(t),$$

$$(3.28c) \quad r(t) = C_{k2}(t)x_k(t) + y(t),$$

$$(3.28d) \quad v(t) = (\mathbb{I}r)(t),$$

in which $\gamma^{-1}\mathbb{I}$ is a causal, strictly contractive operator on $\mathcal{L}^2[0, T]$ and

$$(3.29a) \quad A_k := A + \gamma^{-2}B_1B_1'X_\infty - B_2F_\infty - (B_1D_{21}' + Z_\infty(C_2' + \gamma^{-2}X_\infty B_1D_{21}')) \\ (C_2 + \gamma^{-2}D_{21}B_1'X_\infty),$$

$$(3.29b) \quad \begin{bmatrix} C_{k1} \\ C_{k2} \end{bmatrix} := \begin{bmatrix} F_\infty \\ C_2 + \gamma^{-2}D_{21}B_1'X_\infty \end{bmatrix},$$

$$(3.29c) \quad [B_{k1} \quad B_{k2}] := [-B_1D_{21}' - Z_\infty(C_2' + \gamma^{-2}D_{21}B_1'X_\infty)'] - B_2 - \gamma^{-2}Z_\infty F_\infty',$$

with

$$(3.30a) \quad \dot{Z}_\infty = A_z Z_\infty + Z_\infty A_z' - Z_\infty(C_{k2}'C_{k2} - \gamma^{-2}C_{k1}'C_{k1})Z_\infty + B_1\tilde{D}_\perp'\tilde{D}_\perp B_1', \\ Z_\infty(0) = 0,$$

in which

$$(3.30b) \quad A_z = A - B_1D_{21}'C_2 + \gamma^{-2}B_1\tilde{D}_\perp'\tilde{D}_\perp B_1'X_\infty.$$

Proof. (i) Suppose that a control given by $u = \mathfrak{R}y$ exists such that $\|\mathcal{T}_{zw}\| < \gamma$. Then $u = \mathfrak{R}[C_2 \quad D_{21}]\begin{bmatrix} x \\ w \end{bmatrix}$ together with Theorem 2.3 implies that (3.9) has a solution on $[0, T]$. Since $X_\infty(t)$ exists on $[0, T]$, we may apply the control $u_2 = \mathfrak{R}^*y_2$ to (3.26) to obtain $\|y_1\|_2 < \gamma\|u_1\|_2$ for all $u_1 \neq 0$ by Lemma 3.4 and (3.26). We may now use $u_2 = \mathfrak{R}^*[B_2' \quad I]\begin{bmatrix} p \\ u_1 \end{bmatrix}$ together with the “only if” part of Theorem 3.3 to establish the existence of $Z_\infty(t)$ on $[0, T]$. Sufficiency is immediate from Lemma 3.4 and Theorem 3.3.

(ii) The set of all closed-loop operators corresponding to (3.27) and $u = \mathfrak{R}y$ is the same as the set of all closed-loop operators corresponding to (3.26) and $u_2 = \mathfrak{R}^*y_2$ by Lemma 3.4. Since (3.26) is a problem of the second kind, we may invoke Theorem 3.3 to complete the proof. \square

In the last part of this section, we replace Z_∞ with Y_∞ , which is the solution to the RDE introduced in Theorem 3.3. As we will show, Y_∞ is the dual of X_∞ , and there is a close connection between X_∞ , Y_∞ , and Z_∞ . Before supplying this connection, we give a more general version of the connection between a Riccati equation and the Hamiltonian matrix associated with the two-point boundary value problem.

LEMMA 3.6. *The RDE*

$$(3.31) \quad \dot{P} = A'P + PA - PDP + Q, \quad P(0) = M$$

has a solution on $[0, T]$ if and only if there exists an X such that the boundary value problem

$$(3.32) \quad \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} X - \begin{bmatrix} A & -D \\ -Q & -A' \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} = \frac{d}{dt} \begin{bmatrix} P_1 \\ P_2 \end{bmatrix}$$

has a solution on $[0, T]$ with $P_1(t)$ nonsingular for all $t \in [0, T]$ and $P_2(0)P_1(0)^{-1} = M$. In this case, $P(t) = P_2(t)P_1(t)^{-1}$ is a solution to (3.31) and

$$(3.33) \quad \begin{bmatrix} I \\ P \end{bmatrix} (A - DP) - \begin{bmatrix} A & -D \\ -Q & -A' \end{bmatrix} \begin{bmatrix} I \\ P \end{bmatrix} = \begin{bmatrix} 0 \\ \dot{P} \end{bmatrix}.$$

Proof. Suppose that the RDE has a solution $P(t)$. Then (3.33) is immediate, and the result follows.

Conversely, with $P = P_2P_1^{-1}$ we have from (3.32),

$$\begin{aligned} [P \quad -I] \begin{bmatrix} A & -D \\ -Q & -A' \end{bmatrix} \begin{bmatrix} I \\ P \end{bmatrix} &= [P \quad -I] \left\{ \begin{bmatrix} I \\ P \end{bmatrix} P_1 X - \frac{d}{dt} \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} \right\} P_1^{-1} \\ &= \dot{P}_2 P_1^{-1} - P_2 P_1^{-1} \dot{P}_1 P_1^{-1} = \dot{P}. \quad \square \end{aligned}$$

THEOREM 3.7. *Suppose that (3.9) has a solution $X_\infty(t)$ on $[0, T]$. Then (3.30) has a solution $Z_\infty(t)$ on $[0, T]$ if and only if (3.16) has a solution $Y_\infty(t)$ on $[0, T]$ and $\rho(X_\infty Y_\infty)(t) < \gamma^2$ for all $t \in [0, T]$. Furthermore, $Z_\infty(t) = Y_\infty(I - \gamma^{-2}X_\infty Y_\infty)^{-1}(t)$.*

Proof. A straightforward calculation using (3.9) shows that

$$(3.34) \quad \begin{bmatrix} I & -\gamma^{-2}X_\infty \\ 0 & I \end{bmatrix} H_Y \begin{bmatrix} I & \gamma^{-2}X_\infty \\ 0 & I \end{bmatrix} = H_Z + \begin{bmatrix} 0 & \gamma^{-2}\dot{X}_\infty \\ 0 & 0 \end{bmatrix},$$

where H_Y and H_Z are the Hamiltonian matrices from the boundary value problems associated with the RDEs for Y_∞ and Z_∞ , respectively.

Suppose that Z_∞ exists. Since $X_\infty, Z_\infty \geq 0$, it follows that $(I + \gamma^{-2}X_\infty Z_\infty)(t)$ is nonsingular for all $t \in [0, T]$. Also, $Z_\infty(I + \gamma^{-2}X_\infty Z_\infty)^{-1}(0) = 0$. Using (3.30a) and (3.34), we see that

$$(3.35) \quad H_Y \begin{bmatrix} I + \gamma^{-2}X_\infty Z_\infty \\ Z_\infty \end{bmatrix} = \begin{bmatrix} I + \gamma^{-2}X_\infty Z_\infty \\ Z_\infty \end{bmatrix} (A_z - Z_\infty(C'_{k2}C_{k2} - \gamma^{-2}C'_{k1}C_{k1}))' - \frac{d}{dt} \begin{bmatrix} I + \gamma^{-2}X_\infty Z_\infty \\ Z_\infty \end{bmatrix}.$$

So $Y_\infty = Z_\infty(I + \gamma^{-2}X_\infty Z_\infty)^{-1}$ is a solution to (3.16) by Lemma 3.6. Furthermore,

$$(3.36) \quad \rho[X_\infty Y_\infty] = \rho[X_\infty Z_\infty(I + \gamma^{-2}X_\infty Z_\infty)^{-1}] = \gamma^2 \frac{\rho(X_\infty Z_\infty)}{\gamma^2 + \rho(X_\infty Z_\infty)} < \gamma^2.$$

Conversely, if Y_∞ exists and $\rho(X_\infty Y_\infty) < \gamma^2$, then $I - \gamma^{-2}X_\infty Y_\infty$ is nonsingular on $[0, T]$, $Y_\infty(I - \gamma^{-2}X_\infty Y_\infty)^{-1}(0) = 0$, and from (3.16) we have that

$$\begin{aligned} H_Z \begin{bmatrix} I - \gamma^{-2}X_\infty Y_\infty \\ Y_\infty \end{bmatrix} &= \begin{bmatrix} I - \gamma^{-2}X_\infty Y_\infty \\ Y_\infty \end{bmatrix} (A - B_1 D'_{21} C_2 - Y_\infty(C'_2 C_2 - \gamma^{-2}C'_1 C_1))' \\ &\quad - \frac{d}{dt} \begin{bmatrix} I - \gamma^{-2}X_\infty Y_\infty \\ Y_\infty \end{bmatrix}. \end{aligned}$$

Thus $Z_\infty = Y_\infty(I - \gamma^{-2}X_\infty Y_\infty)^{-1}$ is a solution to (3.30a). \square

Substituting $Z_\infty = Y_\infty(I - \gamma^{-2}X_\infty Y_\infty)^{-1} = (I - \gamma^{-2}Y_\infty X_\infty)^{-1}Y_\infty$ into Theorem 3.5 gives our final result.

THEOREM 3.8. *Given*

$$(3.37a) \quad \dot{x}(t) = A(t)x(t) + B_1(t)w(t) + B_2(t)u(t), \quad x(0) = 0,$$

$$(3.37b) \quad z(t) = C_1(t)x(t) + D_{12}(t)u(t),$$

$$(3.37c) \quad y(t) = C_2(t)x(t) + D_{21}(t)w(t),$$

then (i) an output feedback $u = \Re y$ exists such that $\|\mathcal{T}_{zw}\| < \gamma$ if and only if (a) the Riccati equation (3.9) has a solution X_∞ on $[0, T]$, (b) the Riccati equation (3.16) has a solution Y_∞ on $[0, T]$, and (c) $\rho(X_\infty Y_\infty) < \gamma^2$ for all $t \in [0, T]$; and (ii) every closed-loop operator \mathcal{T}_{zw} with $\|\mathcal{T}_{zw}\| < \gamma$, corresponding to an output feedback control $u = \Re y$, is generated by

$$(3.38a) \quad \dot{x}_k(t) = A_k(t)x_k(t) + B_{k1}(t)y(t) + B_{k2}(t)v(t), \quad x_k(0) = 0,$$

$$(3.38b) \quad u(t) = C_{k1}(t)x_k(t) + v(t),$$

$$(3.38c) \quad r(t) = C_{k2}(t)x_k(t) + y(t),$$

$$(3.38d) \quad v(t) = (\mathbb{U}r)(t),$$

in which $\gamma^{-1}\mathbb{U}$ is a causal, strictly contractive operator on $\mathcal{L}^2[0, T]$ and

$$(3.39a) \quad \begin{bmatrix} C_{k1} \\ C_{k2} \end{bmatrix} := \begin{bmatrix} F_\infty \\ C_2 + \gamma^{-2}D_{21}B_1'X_\infty \end{bmatrix},$$

$$(3.39b) \quad [B_{k1} \ B_{k2}] := -(I - \gamma^{-2}Y_\infty X_\infty)^{-1}[H_\infty' | B_2 + \gamma^{-2}Y_\infty C_1' D_{12}],$$

$$(3.39c) \quad A_k := A + \gamma^{-2}B_1B_1'X_\infty - B_2F_\infty + B_{k1}C_{k2}.$$

Proof. This follows from Theorems 3.5 and 3.7. \square

As a check, we note that (3.38) and (3.39) reduce to those in [9], [11], [13], and [21] in the time-invariant case. In the infinite-horizon case, we must establish an internal stability property; this has already been done in the case of time-invariant problems [9], [11], [13], [21].

Appendix. Suppose that the design problem is described by the equation

$$(A.1) \quad \begin{matrix} n \\ p \\ q \end{matrix} \begin{bmatrix} \dot{x} \\ z \\ y \end{bmatrix} = \begin{bmatrix} A & B_1 & B_2 \\ C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{22} \end{bmatrix} \begin{bmatrix} x \\ w \\ u \end{bmatrix},$$

in which each of the matrices has entries that are continuous functions of time. We will also assume that D_{12} and D_{21} have, respectively, full column and row rank for all $t \in [0, T]$. The controller measures y and generates the control u via $u = \Re y$ and has the task of minimizing the worst-case energy gain between z and w .

It is the aim of this appendix to establish that, without loss of generality, we may assume that

$$(A.2) \quad (i) \quad D_{12}'D_{12} = I_m,$$

$$(A.3) \quad (ii) \quad D_{21}D_{21}' = I_q,$$

$$(A.4) \quad (iii) \quad D_{11} = 0,$$

$$(A.5) \quad (iv) \quad D_{22} = 0.$$

If one prefers, it is possible to have (i) and (ii) replaced by

$$(A.2)' \quad (i)' \quad D'_{12} = [I_m \quad 0_{p-m}],$$

$$(A.3)' \quad (ii)' \quad D_{21} = [I_q \quad 0_{i-q}],$$

which are clearly special cases of (i) and (ii).

To begin, we consider the closed-loop configuration in Fig. A.1, which contains the direct feedthrough matrix from (A.1), the controller, and four scaling matrices. S_1 and S_2 will be chosen to enforce (A.2) and (A.3), while S_3 and S_4 ensure that (A.2)' and (A.3)' are satisfied (should we wish to include them).

Suppose that

$$(A.6) \quad D'_{12} D_{12} = N' N$$

and

$$(A.7) \quad D_{21} D'_{21} = M M'$$

are the Cholesky factorizations. Then the entries of M and N are continuous, since they are expressible in terms of the entries of D_{12} and D_{21} [12, p. 88]. It is easy to check that $S_1 = M^{-1}$ and $S_2 = N^{-1}$ will achieve the desired orthogonalization of D_{21} and D_{12} , and that these scaling matrices will not destroy the continuity properties assumed for the original problem. It is evident from Fig. A.1 that we would design the controller \tilde{K} for the scaled problem, and then back substitute through S_1 and S_2 for K . Next, we introduce the orthogonal extensions D_\perp and \tilde{D}_\perp such that $[D_{12} \quad D_\perp]$ and $[D'_{21} \quad \tilde{D}'_\perp]$ are orthogonal. It is a consequence of Doležal's theorem that the extensions D_\perp and \tilde{D}_\perp exist and are continuous [29, Cor. 3, p. 70]. Setting

$$(A.8) \quad S'_3 = [D_{12} \quad D_\perp], \quad S'_4 = [D'_{21} \quad \tilde{D}'_\perp]$$

in Fig. A.1 leads to a direct feedthrough matrix of the form

$$(A.9) \quad \begin{bmatrix} D_{1111} & D_{1112} & I_m \\ D_{1121} & D_{1122} & 0_{p-m} \\ I_q & 0_{l-q} & D_{22} \end{bmatrix}.$$

The partitioning of D_{11} into D_{11ij} $i, j = 1, 2$ is induced by (A.2)' and (A.3)'. Since S_3 and S_4 are orthogonal and thus norm preserving, these scale factors do not change the set of admissible controllers and therefore the controller representation formula.

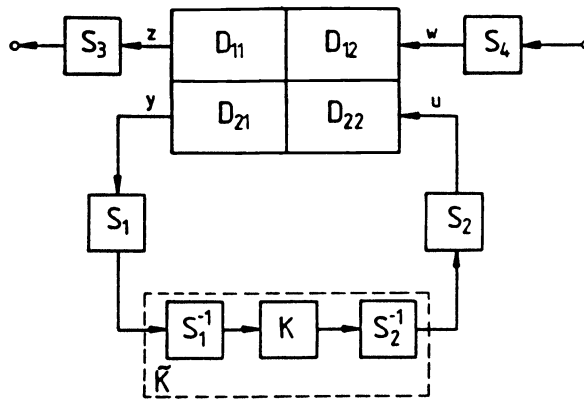


FIG. A.1. Scaling the D -matrix.

We are now in a position to invoke a loop-shifting argument that eliminates both D_{11} and D_{22} , and leads to a significant simplification in the main analysis [11], [24]. Defining

$$(A.10) \quad \gamma_p = \max \left(\sup_{t \in [0, T]} \|D_{1121} \quad D_{1122}\|_2, \sup_{t \in [0, T]} \|D'_{1112} \quad D'_{1122}\|_2 \right),$$

$$(A.11) \quad Q_\infty = -(D_{1111} + D_{1112}(\gamma_p^2 I - D'_{1122}D_{1122})^{-1}D'_{1122}D_{1121}),$$

$$(A.12) \quad F_\infty = (I + Q_\infty D_{22})^{-1}Q_\infty,$$

allows us to mimic the arguments in [11], [24], which remove D_{11} and D_{22} . In conclusion, we mention that the Cholesky factors in the Julia operator of [11] always exist and have continuous entries. \square

Addendum. After the completion of the first version of this paper, we became aware of [5] and [28]. Reference [28] discusses the connections between \mathcal{H}^∞ control and game theory, while [5] applies discrete game theory to the state-feedback \mathcal{H}^∞ -control problem.

REFERENCES

- [1] B. D. O. ANDERSON AND J. B. MOORE, *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, NJ, 1971.
- [2] J. A. BALL AND N. COHEN, *Sensitivity minimization in the \mathcal{H}^∞ norm: Parameterization of all sub-optimal solutions*, Internat. J. Control, 46 (1987), pp. 785–816.
- [3] M. BANKER, *Linear stationary quadratic games*, in Proc. CDC, 1973, pp. 193–197.
- [4] T. BASAR AND G. J. OLSDER, *Dynamic Non-Cooperative Game Theory*, Math. Sci. Engrg., Vol. 160, Academic Press, New York, 1982.
- [5] T. BASAR, *A dynamic game approach to controller design: Disturbance rejection in discrete time*, in Proc. IEEE CDC, Tampa, FL, 1989, pp. 407–414.
- [6] L. D. BERKOWITZ, *A variational approach to differential games*, Adv. Game Theory, in Ann. Math. Study, No. 52, Princeton University Press, Princeton, NJ, 1964, pp. 127–173.
- [7] D. S. BERNSTEIN AND W. M. HADDAD, *LQG control with an \mathcal{H}^∞ performance bound: A Riccati equation approach*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 293–305.
- [8] R. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.
- [9] J. C. DOYLE, K. GLOVER, P. KHARGONEKAR, AND B. FRANCIS, *State-space solutions to standard H^2 and H^∞ control problems*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 831–847.
- [10] 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.
- [11] K. GLOVER, D. J. N. LIMEBEER, J. DOYLE, E. KASENALLY, AND M. G. SAFONOV, *A characterization of all solutions to the four block general distance problem*, SIAM J. Control Optim., 29 (1991), pp. 283–324.
- [12] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, North Oxford Academic, Oxford, UK, 1983.
- [13] M. GREEN, K. GLOVER, D. J. N. LIMEBEER, AND J. C. DOYLE, *A J -spectral factorization approach to H^∞ control*, SIAM J. Control Optim., 28 (1990), pp. 1350–1371.
- [14] P. R. HALMOS, *Introduction to Hilbert Space*, Chelsea, New York, 1951.
- [15] Y. C. HO, A. E. BRYSON, AND S. BARON, *Differential games and optimal pursuit-evasion strategies*, IEEE Trans. Automat. Control, AC-10 (1965), pp. 385–389.
- [16] Y. C. HO, *Differential games, dynamic optimization and generalized control theory*, J. Optim. Theory Appl., 6 (1970), pp. 179–209.
- [17] D. H. JACOBSON, *Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games*, IEEE Trans. Automat. Control, AC-18 (1973), pp. 124–131.
- [18] P. P. KHARGONEKAR, I. R. PETERSEN, AND M. A. ROTEA, *H^∞ optimal control with state feedback*, IEEE Trans. Automat. Control, AC-33 (1988), pp. 783–786.
- [19] P. P. KHARGONEKAR, I. R. PETERSEN, AND K. ZHOU, *Robust stabilization of uncertain systems and H^∞ optimal control*, Tech. Report 87-KPZ, Department of Electrical Engineering, University of Minneapolis, MN; an abridged version appears in IEEE Trans. Automat. Control, AC-35 (1990).

- [20] H. KIMURA AND R. KAWATANI, *Synthesis of H^∞ controllers based on conjugation*, in Proc. IEEE CDC, Austin, TX, 1988.
- [21] D. J. N. LIMEBEER, E. M. KASENALLY, E. JAIMOUKA, AND M. G. SAFONOV, *A characterization of all solutions to the four block general distance problem*, in Proc. IEEE CDC, Austin, TX, Vol. 1, 1988, pp. 878–880.
- [22] I. R. PETERSEN AND D. J. CLEMENTS, *J-spectral factorization and Riccati equations in problems of H^∞ optimization via state feedback*, preprint, 1988.
- [23] I. R. PETERSEN, *Disturbance attenuation and H^∞ optimization: A design method based on the algebraic Riccati equation*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 427–429.
- [24] M. G. SAFONOV AND D. J. N. LIMEBEER, *Simplifying the H^∞ theory via loop shifting*, in Proc. IEEE CDC, Austin, TX, Vol. 2, 1988, pp. 1399–1404.
- [25] W. E. SCHMITTENDORF AND S. J. CITON, *A conjugate point condition for a class of differential games*, J. Optim. Theory Appl., 4 (1969), pp. 109–121.
- [26] W. E. SCHMITTENDORF, *Differential games with open-loop saddle point conditions*, IEEE Trans. Automat. Control, AC-15 (1970), pp. 320–325.
- [27] G. TADMOR, *H^∞ in the time domain: The standard four block problem*, Math. Control Syst. Signal Processing, 3 (1990), pp. 301–324.
- [28] S. WEILAND, *Linear quadratic games, \mathcal{H}^∞ , and the Riccati equation*, Workshop on the Riccati Equation in Control, Systems and Signals, Como, Italy, 1989.
- [29] L. WEISS AND P. L. FALB, *Dolèzal's theorem, linear algebra with continuously parameterized elements, and time varying systems*, Math. Systems Theory, 3 (1969), pp. 67–75.
- [30] P. WHITTLE, *Risk sensitive linear quadratic Gaussian control*, Adv. Appl. Probab., 13 (1981), pp. 764–777.

STOCHASTIC HAMILTON–JACOBI–BELLMAN EQUATIONS*

SHIGE PENG†

Abstract. This paper studies the following form of nonlinear stochastic partial differential equation:

$$-d\Phi_t = \inf_{v \in U} \left\{ \frac{1}{2} \sum_{i,j} [\sigma\sigma^*]_{ij}(x, v, t) \partial_{x_i x_j} \Phi_t(x) + \sum_i b_i(x, v, t) \partial_{x_i} \Phi_t(x) + L(x, v, t) \right. \\ \left. + \sum_{i,j} \sigma_{ij}(x, v, t) \partial_{x_i} \Psi_{j,t}(x) \right\} dt - \Psi_t(x) dW_t, \quad \Phi_T(x) = h(x),$$

where the coefficients σ_{ij} , b_i , L , and the final datum h may be random. The problem is to find an adapted pair $(\Phi, \Psi)(x, t)$ uniquely solving the equation. The classical Hamilton–Jacobi–Bellman (HJB) equation can be regarded as a special case of the above problem. An existence and uniqueness theorem is obtained for the case where σ does not contain the control variable v . An optimal control interpretation is given. The linear quadratic case is discussed as well.

Key words. stochastic control, dynamic programming, Riccati equation, backward stochastic differential equation, stochastic partial differential equation

AMS(MOS) subject classifications. 93E, 60H, 35K

1. Introduction. In this paper, we study a class of backward stochastic partial differential equations which can be derived from certain stochastic optimal control problems where the coefficients may be random variables. We call these kinds of equations stochastic Hamilton–Jacobi–Bellman equations, or HJB equations.

It is well known that the classical HJB equation is the following form of (deterministic) second-order, nonlinear, partial differential equation of parabolic type:

$$(1.1) \quad \begin{aligned} -\partial_t \Phi &= H_1(D^2\Phi, D\Phi, x, t), \\ \Phi(x, T) &= h(x), \end{aligned}$$

where $H_1: \mathbb{R}^{n \times n} \times \mathbb{R}^n \times \mathbb{R}^n \times [0, T] \rightarrow \mathbb{R}$ is given by

$$H_1(A, p, x, v, t) = \inf_{v \in U} \left\{ \frac{1}{2} \text{tr} [\sigma\sigma^*(x, v, t)A] + (b(x, v, t), v) + L(x, v, t) \right\}$$

with

$$\begin{aligned} U &\subset \mathbb{R}^k, \\ \sigma(x, v, t) &: \mathbb{R}^n \times U \times [0, T] \rightarrow \mathbb{R}^n, \\ b(x, v, t) &: \mathbb{R}^n \times U \times [0, T] \rightarrow \mathbb{R}^n, \\ L(x, v, t) &: \mathbb{R}^n \times U \times [0, T] \rightarrow \mathbb{R}, \\ h(x) &: \mathbb{R}^n \rightarrow \mathbb{R}. \end{aligned}$$

This equation has a stochastic control interpretation: Let $\{W_t, t \geq 0\}$ be, for example, a one-dimensional standard Wiener process. For any given initial data $(x, t) \in \mathbb{R}^n \times [0, T]$, consider the following stochastic control system:

$$(1.2) \quad \begin{aligned} dz_s &= b(z_s, v_s, s) ds + \sigma(z_s, v_s, s) dW_s, \quad t \leq s \leq T, \\ z_t &= x, \end{aligned}$$

* Received by the editors January 29, 1990; accepted for publication (in revised form) January 3, 1991. This work was partially supported by State Education Committee (SEDC) Foundation for Young Academics, the People's Republic of China, and by National Natural Science Foundation of China.

† Department of Mathematics, Shandong University, Jinan 250100, China.

where v_s , $0 \leq s \leq T$ is a U -valued adapted stochastic process called admissible control. The optimal control problem is to minimize the following functional:

$$(1.3) \quad J_{x,t}(v) = E \left\{ \int_t^T L(z_s, v_s, s) ds + h(z_T) \right\}$$

over admissible controls. We can define the following value function:

$$\Phi(x, t) = \inf_v J_{x,t}(v),$$

which is defined on $\mathbb{R}^n \times [0, T]$ with values in \mathbb{R} . It can be proved that, under some reasonable assumptions, the value function is the unique solution to (1.1). We refer to Fleming and Rishel [6], Bensoussan and Lions [3], Krylov [11], [12], and Lions [14] for details.

In (1.1), the functions $\sigma(x, v, t)$, $b(x, v, t)$, $L(x, v, t)$, and $h(x)$ are all deterministic. The objective of this paper is to study the case when σ, b, L, h are possibly randomly perturbed. A typical case is when $h = h(x, \omega)$ is \mathcal{F}_T measurable. Here \mathcal{F}_t is the filtration generated by the Brownian motion W_t . In this case, the corresponding value function becomes random. In fact, it is a solution of the following type of backward stochastic partial differential equation:

$$(1.4) \quad \begin{aligned} -d\Phi_t &= H(D^2\Phi, D\Phi, D\Psi, x, t) dt - \Psi dW_t, \\ \Phi(x, T) &= h(x), \end{aligned}$$

with

$$(1.5) \quad H(A, p, x, B, t) = \inf_{v \in U} \{ \frac{1}{2} \text{tr}[\sigma \sigma^*(x, v, t) A] + (b(x, v, t), v) + (\sigma(x, v, t), B) + L(x, v, t) \}.$$

Here for any fixed $x \in \mathbb{R}^n$, $\Phi_t(x, \omega)$ and $\Psi_t(x, \omega)$ are \mathcal{F}_t -adapted processes. We call this equation a stochastic HJB equation.

The main result in this paper asserts the existence and uniqueness of an adapted pair (Φ, Ψ) with values in $H^1(\mathbb{R}^n) \times L^2(\mathbb{R}^n)$ satisfying (1.4). Due to the limitations of our approach, we can only deal with the case where σ does not contain the control variable v . Some nondegeneracy assumption is also needed.

The uniqueness part follows by an application of the backward Itô rule, while the existence is obtained by solving the equation iteratively and applying Itô's rule to estimate the errors between successive estimates.

The unusual feature here is the fact that Ψ is uniquely determined although no time derivative is specified. For understanding this point, let us recall the finite-dimensional situation (see Theorem 2.1 below) in which (1.4) is replaced by

$$(1.6) \quad X = x_t + \int_t^T b(x_s, m_s, s) ds + \int_t^T [\sigma(x_s) + m_s] dW_s, \quad 0 \leq t \leq T,$$

where the terminal value X is \mathcal{F}_T -measurable. Applying Itô's rule to the difference of two solutions $\rho_t = E(|x_t - x'_t|^2)$ yields

$$\begin{aligned} \rho_t + E \int_t^T |\sigma(x) - \sigma(x') + m - m'|^2 ds \\ = 2E \int_t^T \langle b(x, m) - b(x', m'), x - x' \rangle ds, \end{aligned}$$

$$0 \leq t \leq T.$$

Simple estimation then yields

$$\rho_t + \frac{1}{2}E \int_t^T |m - m'|^2 \leq C \int_t^T \rho_s ds.$$

Now ignoring the m -term and using Gronwall's inequality yields $x = x'$, and hence $m = m'$. The same method is used to establish convergence of Picard iteration of solution (x, m) . With more work the same idea lies at the basis of the existence and uniqueness problem for (1.4).

In the paper, a connection of (1.4) with an optimal control problem is described. This connection is similar to that of (1.1) with the optimal control problems (1.2) and (1.3). It should be pointed out that the function H given in (1.5) is itself allowed to be an adapted process.

As an important special case, we study the linear quadratic optimal control problem with random coefficients. A kind of matrix-valued backward stochastic differential equation, called a stochastic Riccati equation, is investigated, and a result of Bismut [4] is nontrivially generalized.

The attention to such kind of backward stochastic differential equations (SDEs) determined by a pair of unknown adapted processes was originated in the study of the stochastic maximum principle for optimal control systems, where the adjoint processes are a pair of adapted processes. This adapted pair can be characterized by a linear backward stochastic differential equation called an adjoint equation (see Bensoussan [1], Bismut [5], Haussmann [7], Kushner [13], and Peng [17], for the finite-dimensional case, and Bensoussan [2] and Hu and Peng [9] for the infinite-dimensional case). The study of forward stochastic partial differential equations can be found in Pardoux [15]. Recently, Pardoux and Peng [16] obtained an existence and uniqueness result for nonlinear backward stochastic differential equations. The infinite-dimensional case (for semi linear evolution systems) can be found in Hu and Peng [8]. The last two results have been applied in this paper. Some results of this paper were briefly announced in Peng [18].

This paper is organized as follows. In the next section, we recall some preliminary results, mainly concerning existence and uniqueness theorems for the backward stochastic differential equations. Section 3 is devoted to the stochastic control interpretation of the stochastic HJB equations and a related verification theorem. The existence and uniqueness result is discussed in § 4. In § 5, we consider linear quadratic optimal control problems with random coefficients.

2. Preliminaries. We first recall some results concerning adapted backward stochastic differential equations and a generalized Itô formula that will be used in the following.

2.1. Finite-dimensional case. Let (Ω, \mathcal{F}, P) be a probability space equipped with filtration \mathcal{F}_t^* . Let $\{W_t, t \geq 0\}$ be a d -dimensional standard Wiener process defined on it. We denote

$$\mathcal{F}_t = \sigma\{W_s; 0 \leq s \leq t\}.$$

For any given Hilbert (or Euclidean) space H , we will denote by $\mathcal{M}^2(H)$ the space of all \mathcal{F}_t -adapted processes with values in H , such that

$$E \int_0^T |x_t|_H^2 dt < \infty, \quad \forall x \in \mathcal{M}^2(H).$$

Obviously, $\mathcal{M}^2(H)$ is a Hilbert space. We denote by (\cdot, \cdot) (respectively, $|\cdot|$), the scalar product (respectively, norm) of a Euclidean space. Note that the space $\mathcal{L}(H; \mathbb{R}^n)$ is also a Hilbert (respectively, Euclidean) space under the scalar product

$$(m_1, m_2) = \text{tr}(m_1^* m_2), \quad \forall m_1, m_2 \in \mathcal{L}(\mathbb{R}^n; H),$$

where we denote by $\text{tr}(A)$, the trace of an $n \times n$ matrix A . Let the following functions be given:

$$\begin{aligned} f(x, m, t, \omega) &: \mathbb{R}^n \times \mathcal{L}(\mathbb{R}^d; \mathbb{R}^n) \times [0, T] \times \Omega \rightarrow \mathbb{R}^n, \\ \sigma(x, m, t, \omega) &: \mathbb{R}^n \times \mathcal{L}(\mathbb{R}^d; \mathbb{R}^n) \times [0, T] \times \Omega \rightarrow \mathcal{L}(\mathbb{R}^d; \mathbb{R}^n), \\ X(\omega) &: \Omega \rightarrow \mathbb{R}^n. \end{aligned}$$

We assume

- (i) for each $(x, m) \in \mathbb{R}^n \times \mathcal{L}(\mathbb{R}^d; \mathbb{R}^n)$, $f(x, m, \cdot) \in \mathcal{M}^2(\mathbb{R}^n)$
 $\sigma(x, m, \cdot) \in \mathcal{M}^2(\mathcal{L}(\mathbb{R}^d; \mathbb{R}^n))$;
- (2.1) (ii) for each $(t, \omega) \in [0, T] \times \Omega$, $f(x, m, t, \omega)$, $\sigma(x, m, t, \omega)$ are differentiable with respect to (x, m) , and the derivatives are all bounded;
- (iii) $X(\omega)$ is \mathcal{F}_T -measurable, and $E|X|^2 < \infty$.

Consider the following backward stochastic differential equation:

$$(2.2) \quad X = x_t + \int_t^T f(x_s, m_s, s) ds + \int_t^T [\sigma(x_s) + m_s] dW_s.$$

Our problem is to find a pair of adapted $\mathbb{R}^n \times \mathcal{L}(\mathbb{R}^d; \mathbb{R}^n)$ -valued processes (x_s, m_s) , which solves equation (2.2). We have the following result.

THEOREM 2.1. *Assume (2.1) holds. Then, there exists a unique pair (x_s, m_s) in $\mathcal{M}^2(\mathbb{R}^n) \times \mathcal{M}^2(\mathcal{L}(\mathbb{R}^d; \mathbb{R}^n))$, which solves (2.2). We have*

$$(2.3) \quad E \sup_{t \in [0, T]} |x_t|^2 < \infty.$$

Moreover, if

$$(2.4) \quad K = \sup_{\omega} \left\{ |X|^2 + \int_0^T |f(0, 0, s)|^2 ds + \int_0^T |\sigma\sigma^*(0, s)|^2 ds \right\} < \infty,$$

then

$$(2.5) \quad |x_t|^2 \leq K e^{C(T-t)},$$

where C is a positive constant depending only on the bound of $|f_x|, |\sigma_x|$.

Proof. The proof of existence and uniqueness for (2.2) and (2.3) can be found in [16]. We only prove (2.5). Applying Itô's formula to (2.2), we have, for $0 \leq r \leq t \leq T$,

$$(2.6) \quad E^{\mathcal{F}_r} |x_t|^2 + E^{\mathcal{F}_r} \int_t^T |(\sigma(x_s, s) + m_s)|^2 ds = E^{\mathcal{F}_r} |X|^2 - 2E^{\mathcal{F}_r} \int_t^T (x_s, f(x_s, m_s, s)) ds.$$

Thus

$$\begin{aligned} E^{\mathcal{F}_r} |x_t|^2 + \frac{1}{2} E^{\mathcal{F}_r} \int_t^T |m_s|^2 ds &\leq E^{\mathcal{F}_r} |X|^2 + E^{\mathcal{F}_r} \int_t^T |\sigma(x_s, s)|^2 ds \\ &\quad + 2E^{\mathcal{F}_r} \int_t^T |x_s| |f(x_s, m_s, s)| ds. \end{aligned}$$

Since the derivatives of f, σ with respect to (x, m) are bounded,

$$\begin{aligned} 2|x||f(x, m, s)| &\leq 2|x||f(0, 0, s)| + 2C_1|x|(|x| + |m|) \\ &\leq |f(0, 0, s)|^2 + (4C_1^2 + 2C_1 + 1)|x|^2 + \frac{1}{4}|m|^2, \\ |\sigma(x, s)|^2 &\leq |\sigma(0, s)|^2 + C_1|x|^2. \end{aligned}$$

It follows that

$$\begin{aligned} E^{\mathcal{F}_r}|x_t|^2 + \frac{1}{4}E^{\mathcal{F}_r} \int_t^T |m_s|^2 ds \\ \leq E^{\mathcal{F}_r}|X|^2 + E^{\mathcal{F}_r} \int_t^T (|f(0, 0, s)|^2 + |\sigma(0, s)|^2 + (4C_1^2 + 3C_1 + 1)|x|^2) ds \\ \leq K + \int_t^T (4C_1^2 + 3C_1 + 1)E^{\mathcal{F}_r}|x|^2 ds. \end{aligned}$$

Applying Gronwall's inequality, we obtain finally

$$E^{\mathcal{F}_r}|x_t|^2 \leq K e^{C(T-t)}, \quad \forall 0 \leq r \leq t \leq T,$$

with $C = 4C_1^2 + 3C_1 + 1$. This implies (2.5). \square

2.2. Infinite-dimensional case. Let V, H be two separable Hilbert spaces such that V is densely embedded in H . We identify H with its dual space, and denote by V' the dual of V . We have then

$$V \subset H \subset V'.$$

We will denote by $\|\cdot\|, |\cdot|, \|\cdot\|_*$ the norms of V, H , and V' , respectively; by $\langle \cdot, \cdot \rangle$ the duality product between V and V' , and by (\cdot, \cdot) the scalar product in H . For any $\varphi \in \mathcal{M}^2(\mathcal{L}(\mathbb{R}^n; H))$, we can define an H -valued random variable, called stochastic integral $\int_0^t \varphi_s dW_s$ by (see [15])

$$\left(h, \int_0^t \varphi_s dW_s \right) = \int_0^t (\varphi_s^* h) dW_s, \quad \forall h \in H.$$

Let the following functions be given:

$$\begin{aligned} A(t, \omega) &: [0, T] \times \Omega \rightarrow \mathcal{L}(V; V'), \\ f(y, t, \omega) &: H \times [0, T] \times \Omega \rightarrow V', \\ g(y, z, t, \omega) &: V \times \mathcal{L}(\mathbb{R}^d; H) \times [0, T] \times \Omega \rightarrow H, \\ \eta(y, t, \omega) &: H \times [0, T] \times \Omega \rightarrow \mathcal{L}(\mathbb{R}^d; H), \\ Y(\omega) &: \Omega \rightarrow H. \end{aligned}$$

We assume the following:

$$\begin{aligned} (2.7) \quad &\text{For each } (y, z), A(t), f(y, t), g(y, z, t), \eta(y, t) \text{ are } \mathcal{F}_t\text{-adapted,} \\ &\text{such that, } f(0, t) \in \mathcal{M}^2(V'), \eta(0, t) \in \mathcal{M}^2(\mathcal{L}(\mathbb{R}^d; H)), \\ &g(0, 0, t) \in \mathcal{M}^2(H), A(t) \in \mathcal{M}^2(\mathcal{L}(V, V')), \\ &\sup_{t, \omega} \|A(t, \omega)\|_{\mathcal{L}(V, V')} \leq C; \end{aligned}$$

$$\begin{aligned} (2.8) \quad &\exists \alpha > 0 \text{ and } \lambda, \text{ such that, } \forall y \in V, \\ &\langle A(t)y, y \rangle + \lambda|y|^2 \geq \alpha\|y\|^2, \quad \forall t; \end{aligned}$$

$$\begin{aligned}
& |g(y_1, z_1, t) - g(y_2, z_2, t)| \leq C(\|y_1 - y_2\| + |z_1 - z_2|), \\
& \mathbf{V}(y_1, z_1), (y_2, z_2) \in (V \times \mathcal{L}(\mathbb{R}^d; H)), \\
(2.9) \quad & |\eta(y_1, t) - \eta(y_2, t)| \leq C|y_1 - y_2|, \quad \mathbf{V}y_1, y_2 \in H, \\
& |\langle f(y_1, t) - f(y_2, t), y \rangle| \leq C|y_1 - y_2| \cdot \|y\|, \quad \mathbf{V}(y_1, y_2) \in H, \quad y \in V.
\end{aligned}$$

Consider the following semilinear backward stochastic equation:

$$(2.10) \quad Y = y_t + \int_t^T (A(s)y_s + f(y_s, s) + g(y_s, z_s, s)) ds + \int_t^T (\eta(y_s, s) + z_s) dW_s.$$

The problem is to find a pair of adapted processes (y, z) satisfying the above equation. We have the following theorem (see [8] for proof).

THEOREM 2.2. *We assume (2.7)–(2.9). Then there exists a unique pair $(y, z) \in \mathcal{M}^2(V) \times \mathcal{M}^2(\mathcal{L}(\mathbb{R}^d; H))$, satisfying the backward evolution equation (2.10).*

Remark. We can also obtain an estimate similar to (2.5).

We need also the following generalized Itô's formula due to Kunita [10].

THEOREM 2.3. *Let $F_t(x)$, $(x, t) \in \mathbb{R}^n \times [0, T]$ be a random field which is continuous in (x, t) almost surely, satisfying*

- (i) *For each t , $F_t(\cdot)$ is a C^2 -map from \mathbb{R}^n into R almost surely;*
- (ii) *For each x , $F_t(x)$ is a continuous semimartingale and it satisfies*

$$F_t(x) = F_0(x) + \sum_{j=1}^m \int_0^t f_s^j(x) dY_s^j, \quad \mathbf{V}x \in \mathbb{R}^n,$$

almost surely, where Y_s^j , $j = 1, \dots, d$ are continuous semimartingales, $f_s^j(x)$, $x \in \mathbb{R}^n$, $s \in [0, T]$, are random fields that are continuous in (x, s) and satisfy

- (a) *for each s , $f_s^j(x)$ is a C^2 -map from \mathbb{R}^n to R ,*
- (b) *for each x , $f_s^j(x)$ is an adapted process.*

Let $X_t = (X_t^1, \dots, X_t^n)$ be continuous semimartingales. Then we have

$$\begin{aligned}
(2.11) \quad F_t(X_t) &= F_0(X_0) + \sum_{j=1}^m \int_0^t f_s^j(X_s) dY_s^j + \sum_{i=1}^n \int_0^t \partial_{x_i} F_s(X_s) dX_s^i \\
&\quad + \sum_{j=1}^m \sum_{i=1}^n \int_0^t \partial_{x_i} f_s^j(X_s) d\langle Y^j, X^i \rangle_s \\
&\quad + \sum_{i,j=1}^n \int_0^t \partial_{x_i x_j} F_s(X_s) d\langle X^i, X^j \rangle_s,
\end{aligned}$$

where $\langle \cdot, \cdot \rangle_s$ stands for the quadratic variation of semimartingales.

3. Optimal control formulation of the stochastic HJB equation. We begin by considering a stochastic optimal control system in which all coefficients may be also stochastic processes. This can lead us to formulate what we call stochastic HJB equations. We present a general form of such equations. The discussion is formal because, in general, we have no sufficient regularity properties of the value function. We will also discuss the so-called “verification theorem” approach in which the rigorous proof can be easily given.

3.1. Optimal control system. Let $\{W_t^1; t \geq 0\}$ be a p -dimensional standard Wiener process in (Ω, \mathcal{F}, P) which is independent of $\{W_t; t \geq 0\}$. Without loss of generality, we consider only the case where W_t is a one-dimensional Brownian motion, i.e., $d = 1$. We assume

$$\mathcal{F}_t^* = \sigma\{W_s, W_s^1; 0 \leq s \leq t\}.$$

Consider the following stochastic control system parametrized by the initial data $(x, t) \in \mathbb{R}^n \times [0, T]$:

$$(3.1) \quad dy_s = b(y_s, v_s, s) ds + \sigma(y_s, v_s, s) dW_s + \pi(y_s, v_s, s) dW_s^1, \quad t \leq s \leq T, \quad y_t = x,$$

where

$$\begin{aligned} b(x, v, t) &: \mathbb{R}^n \times \mathbb{R}^k \times [0, T] \rightarrow \mathbb{R}^n, \\ \sigma(x, v, t) &: \mathbb{R}^n \times \mathbb{R}^k \times [0, T] \rightarrow \mathbb{R}^n, \\ \pi(x, v, t) &: \mathbb{R}^n \times \mathbb{R}^k \times [0, T] \rightarrow \mathcal{L}(\mathbb{R}^p; \mathbb{R}^n). \end{aligned}$$

An admissible control is an \mathcal{F}_t^* -adapted process v_t with values in U , such that

$$E \int_0^T |v_t|^2 dt < \infty.$$

We denote the set of admissible controls by \mathcal{U} . By analogy with the classical optimal control problem, our problem is, for a given initial data $y_0 = x_0$ in (3.1), to find an admissible control v_t which minimizes the following cost function:

$$(3.2) \quad E \left\{ \int_0^T L(y_s, v_s, s) ds + h(y_T) \right\},$$

where we denote

$$(3.3) \quad \begin{aligned} L(x, v, t) &: \mathbb{R}^n \times \mathbb{R}^k \times [0, T] \rightarrow \mathbb{R}, \\ h(x, \omega) &: \mathbb{R}^n \times \Omega \rightarrow \mathbb{R}. \end{aligned}$$

The main difference between the classical optimal control problem and the above one is that h is a random variable. We assume that $h(x)$ is \mathcal{F}_T -measurable.

Remark. We can also treat the case where σ, π, b , and L are \mathcal{F}_t -adapted random processes. The approach and the conclusion are the same.

Following the idea of dynamic programming, for any fixed initial data (x, t) in (3.1), we minimize the following cost function over admissible controls:

$$(3.4) \quad J_{x,t}(v) = E^{\mathcal{F}_t} \int_t^T L(y_s, v_s, s) ds + h(y_T).$$

The value function is defined as follows:

$$\Phi_t(x) = \inf_{v \in \mathcal{U}} J_{x,t}(v).$$

Observe that, for any fixed x , Φ_t is an \mathcal{F}_t -adapted process with values in \mathbb{R} . In general, Φ_t is not a bounded variation function with respect to t . In fact, we can only expect that Φ_t is a semimartingale

$$(3.5) \quad \Phi_t(x) = \int_t^T \Gamma_s(x) ds - \int_t^T \Psi_s(x) dW_s, \quad x \in \mathbb{R}^n, \quad 0 \leq t \leq T,$$

where, for each $x \in \mathbb{R}^n$ and $\Gamma_s(x), \Psi_s(x)$ are \mathcal{F}_s -adapted real processes.

If $\Phi_t(x)$ can be expressed in the form (3.5), and if $\Phi_t(x), \Gamma_t(x), \Psi_t(x)$ are almost surely continuous in (x, t) and are smooth enough with respect to x , then we can prove that the pair $(\Phi_t(x), \Psi_t(x))$ satisfies the following stochastic partial differential equation:

$$(3.6) \quad -d\Phi_t = H(D^2\Phi, D\Phi, D\Psi, x, t) dt - \Psi dW_t, \quad \Phi(x, T) = h(x),$$

or

$$(3.7) \quad \Phi_t = h(x) + \int_t^T H(D^2\Phi, D\Phi, D\Psi, x, s) ds - \int_t^T \Psi dW_s,$$

with

$$H(A, p, x, B, t) = \inf_{v \in U} \{ \text{tr}[a(x, v, t)A] + (b(x, v, t), p) + (\sigma(x, v, t), B) + L(x, v, t) \},$$

$$A \in \mathbb{R}^{n \times n}, \quad B \in \mathbb{R}^n,$$

where $D\Phi$ is the gradient of Φ , $D^2\Phi$ is the Hessian of Φ , and

$$a(x, v, t) = \frac{1}{2}[\sigma\sigma^*(x, v, t) + \pi\pi^*(x, v, t)].$$

Equation (3.6) or (3.7) is called a stochastic HJB equation. Comparing (3.6) with the classical HJB equation, which is a deterministic partial differential equation, an interesting feature of this equation is that its solution consists of a pair (Φ_t, Ψ_t) .

The following procedure can verify that (Φ_t, Ψ_t) as in (3.5) satisfies (3.7). First, similar to the classical case, we introduce Bellman's optimality principle

$$(3.8) \quad \begin{aligned} \Phi_t(x) &= \inf_{v \in \mathcal{U}} E^{\mathcal{F}_t} \left\{ \int_t^{t+r} L(y_s, v_s, s) ds + E^{\mathcal{F}_{t+r}} \left(\int_{t+r}^T L(y_s, v_s, s) ds + h(y_T) \right) \right\} \\ &= \inf_{v \in \mathcal{U}} E^{\mathcal{F}_t} \left\{ \int_t^{t+r} L(y_s, v_s, s) ds + \Phi_{t+r}(y_{t+r}) \right\}. \end{aligned}$$

Then, applying Itô-Kunita's formula (2.11) to the process $\Phi_s(y_s)$ yields

$$\begin{aligned} \Phi_{t+r}(y_{t+r}) &= \Phi_t(x) + E^{\mathcal{F}_t} \int_t^{t+r} [\text{tr}(a(y_s, v_s, s)D^2\Phi_s(y_s)) + (b(y_s, v_s, s), D\Phi_s(y_s)) \\ &\quad + \text{tr}(\sigma(y_s, v_s, s)D\Psi_s(y_s)) + \Gamma_s(y_s)] ds. \end{aligned}$$

This with (3.8) implies

$$\begin{aligned} r^{-1} \inf_{v \in \mathcal{U}} E^{\mathcal{F}_t} \int_t^{t+r} [L(y_s, v_s, s) + \text{tr}(a(y_s, v_s, s)D^2\Phi_s(y_s)) + (b(y_s, v_s, s), D\Phi_s(y_s)) \\ + (\sigma(y_s, v_s, s), D\Psi_s(y_s)) + \Gamma_s(y_s)] ds = 0. \end{aligned}$$

Passing limit as $r \rightarrow 0$, we have

$$\Gamma_t(x) = -H(D^2\Phi_t(x), D\Phi_t(x), D\Psi_t(x), x, t).$$

Substituting it into (3.5) yields (3.7).

Unfortunately, as in the classical case, even when the coefficients b, σ, π, L, h are sufficiently regular with respect to (x, v) , it is still difficult to verify the regularities of $(\Phi_t(x), \Psi_t(x))$.

3.2. Verification theorem approach. We now show that a sufficiently smooth solution of (3.7) coincides with the value function. In the classical case, this approach is called the verification theorem approach (see [6]).

Indeed, let a sufficiently smooth pair $(\varphi_t(x), \psi_t(x))$ be a solution of (3.7). Assume

- (i) For each t , $(\varphi_t(x), \psi_t(x))$ is a C^2 -map from \mathbb{R}^n into $\mathbb{R}^1 \times \mathbb{R}^d$.
- (ii) For each x , $(\varphi_t(x), \psi_t(x))$, $(D\varphi_t(x), D^2\varphi_t(x), D\psi_t(x))$ are continuous \mathcal{F}_t -adapted processes.

Without loss of generality, let $u_t(x)$ be a random field that is $\mathcal{F}_1 \times \mathcal{B}(\mathbb{R}^n)/\mathcal{B}(\mathbb{R}^k)$ -measurable with values in \mathcal{U} , such that

$$\begin{aligned} & \text{tr}[a(x, u_t, t)D^2\varphi_t] + (b(x, u_t, t), D\varphi_t) + (\sigma(x, u_t, t), D\psi_t) + L(x, u_t, t) \\ & = H(D^2\varphi_t(x), D\varphi_t(x), D\psi_t(x), x, t). \end{aligned}$$

Furthermore, suppose that $u_t(x)$ is regular enough such that the “feedback” control system

$$\begin{aligned} dy_s &= b(y_s, u_s(y_s), s) ds + \sigma(y_s, u_s(y_s), s) dW_s + \pi(y_s, u_s(y_s), s) dW_s^1, \\ t &\leq s \leq T, \quad y_t = x, \end{aligned}$$

is well posed. Finally, suppose that $H(D^2\varphi_t(x), D\varphi_t(x), D\psi_t(x), x, t)$ is smooth with respect to (x, t) . Then $(\varphi_t(x), \psi_t(x))$ coincides with $(\Phi_t(x), \Psi_t(x))$. Furthermore, for any initial data $(x_0, 0)$, $u_s^* = u_s(y_s)$ minimize the cost function (3.2).

Indeed, let v_s be an admissible control, and let z_s be the corresponding solution

$$dz_s = b(z_s, v_s, s) ds + \sigma(z_s, v_s, s) dW_s + \pi(z_s, v_s, s) dW_s^1, \quad t \leq s \leq T, \quad z_t = x.$$

From Itô–Kunita’s formula,

$$\begin{aligned} E^{\mathcal{F}_t}\varphi_T(z_T) &= \varphi_t(x) + E^{\mathcal{F}_t} \int_t^T [\text{tr}(a(z_s, v_s, s)D^2\varphi_s(z_s)) + (b(z_s, v_s, s), D\varphi_s(z_s)) \\ &\quad + (\sigma(z_s, v_s, s), D\psi_s(z_s)) - H(D^2\varphi_s(z_s), D\varphi_s(z_s), D\psi_s(z_s), z_s, s)] ds. \end{aligned}$$

This yields

$$\begin{aligned} (3.9) \quad \varphi_t(x) &\leq E^{\mathcal{F}_t} \left\{ \int_t^T L(z_s, v_s, s) ds + h(z_T) \right\} \\ &= J_{x,t}(v). \end{aligned}$$

On the other hand, if we take the above $u_s^* = u_s(y_s)$ as a control, then, again from Itô–Kunita’s formula,

$$\begin{aligned} E^{\mathcal{F}_t}\varphi_T(y_T) &= \varphi_t(x) + E^{\mathcal{F}_t} \int_t^T [\text{tr}(a(y_s, u_s^*, s)D^2\varphi_s(y_s)) + (b(y_s, u_s^*, s), D\varphi_s(y_s)) \\ &\quad + (\sigma(y_s, u_s^*, s), D\psi_s(y_s)) - H(D^2\varphi_s(y_s), D\varphi_s(y_s), D\psi_s(y_s), y_s, s)] ds. \end{aligned}$$

It follows from the definition of u_s^* that

$$\begin{aligned} \varphi_t(x) &= E^{\mathcal{F}_t} \left\{ \int_t^T L(y_s, u_s^*, s) ds + h(y_T) \right\} \\ &= J_{x,t}(u^*). \end{aligned}$$

This with (3.9) implies

$$\varphi_t(x) = \inf_{v \in \mathcal{U}} J_{x,t}(v) = \Phi_t(x).$$

Consequently, $\psi_t(x) = \Psi_t(x)$. The last assertion is easy to see since

$$E \left\{ \int_0^T L(y_s, v_s, s) ds + h(y_T) \right\} = EE^{\mathcal{F}_0} \left\{ \int_0^T L(y_s, v_s, s) ds + h(y_T) \right\}.$$

4. Existence and uniqueness. An existence and uniqueness theorem for the stochastic HJB equation (3.7), in terms of Sobolev space technique, is given in this section. We can only treat the case where no control variable appears in the diffusion coefficients σ and π . We also need a nondegeneracy assumption. The corresponding control system (3.1) becomes

$$(4.1) \quad dy_s = b(y_s, v_s, s) ds + \sigma(y_s, s) dW_s + \pi(y_s, s) dW_s^1, \quad t \leq s \leq T, \quad y_t = x,$$

with the following cost function to be minimized:

$$(4.2) \quad J_{x,t}(v) = E^{\mathcal{F}_t} \left\{ \int_t^T L(y_s, v_s, s) ds + h(y_T) \right\}.$$

We assume that

- (i) For each t , $b(x, v, t)$, $\sigma(x, t)$, $\pi(x, t)$, $L(x, v, t)$, $h(x)$ are continuous in (x, v) ;
- (4.3) (ii) For each (v, t) , they are continuously differentiable in x ;
- (iii) They and all their derivatives in x are bounded;
- (iv) For each (x, v) , they are continuous in t .

We also need the following nondegeneracy assumption on π :

$$(4.4) \quad \pi\pi^*(x, t) \geq 2\alpha I, \quad \forall (x, t),$$

where α is a positive constant.

Remark. Technically, the above nondegeneracy assumption will be used to satisfy (2.8). In general, if we have no such kind of condition, some kind of notion like “viscosity solution” seems necessary. A typical case is when h is deterministic and $\sigma = \pi = 0$. In this case the most suitable notation is the viscosity solution (see [14]).

Our problem is to find a unique pair $(\Phi_t(x), \Psi_t(x))$ satisfying (3.7), where the function H becomes

$$\begin{aligned} a(x, t) &= \frac{1}{2}(\sigma\sigma^*(x, t) + \pi\pi^*(x, t)), \\ H(A, p, B, x, t) &= \text{tr}(a(x, t)A) + (\sigma(x, t), B) + \inf_{v \in U} \{(b(x, v, t), p) + L(x, v, t)\} \\ &\quad \forall x \in \mathbb{R}^n, \quad t \in [0, T], \quad A \in S^n, \quad B \in \mathbb{R}^n. \end{aligned}$$

We introduce the following Sobolev space:

$$H^1(\mathbb{R}^n) = \{u \in L^2(\mathbb{R}^n): Du \in L^2(\mathbb{R}^n)\}.$$

We identify $L^2(\mathbb{R}^n)$ with its dual space, and denote the dual space of $H^1(\mathbb{R}^n)$ by $H^{-1}(\mathbb{R}^n)$. This gives us a triple as in § 2.2

$$(H^1(\mathbb{R}^n), L^2(\mathbb{R}^n), H^{-1}(\mathbb{R}^n)) = (V, H, V').$$

We can make the following assertion.

THEOREM 4.1. *Assume that (4.3) and (4.4) hold. Then there exists a unique pair $(\Phi_t(x), \Psi_t(x))$ in $(\mathcal{M}^2(V), \mathcal{M}^2(H))$ satisfying the stochastic HJB equation (3.7).*

To prove this theorem, we need a lemma that generalizes Theorem 2.2. Let $g(y, z, t)$, Y , $A(t)$ be as in Theorem 2.2. Let

$$\begin{aligned} A_1(t, \omega) &: [0, T] \times \Omega \rightarrow \mathcal{L}(V; V'), \\ f(y, z, t, \omega) &: H \times H \times [0, T] \times \Omega \rightarrow V'. \end{aligned}$$

We assume that the following are true:

(4.5) For each (y, z) , $A_1(t)$ and $f(y, z, t)$ are \mathcal{F}_t -adapted, $f(0, 0, t) \in \mathcal{M}^2(V)$;

(4.6) $\sup_{t, \omega} \|A_1(t, \omega)\|_{\mathcal{L}(V, V)} \leq C, \quad \langle A_1(t)y, y \rangle \geq 0, \quad \forall y \in V;$

(4.7) $|\langle f(y_1, z_1, t) - f(y_2, z_2, t), y \rangle| \leq C|y_1 - y_2|_H|y|_V + |z_1 - z_2|_H \langle 2A_1(t)y, y \rangle^{1/2},$
for each $(y_1, z_1), (y_2, z_2) \in H \times H$, for each $y \in V$.

Consider the following semilinear backward stochastic equation:

$$(4.8) \quad Y = y_t + \int_t^T (A(s)y_s + A_1(s)y_s + f(y_s, z_s, s) + g(y_s, z_s, s)) ds + \int_t^T z_s dW_s.$$

We have the following lemma.

LEMMA 4.2. Let $g(y, z, t)$, Y , $A(t)$ satisfy the assumptions of Theorem 2.2. Let (4.5)–(4.7) hold. Then, there exists a unique pair $(y, z) \in \mathcal{M}^2(V) \times \mathcal{M}^2(H)$, satisfying the backward evolution equation (4.8).

Proof. (i). *Uniqueness.* Let $(y_s^1, z_s^1), (y_s^2, z_s^2)$ be two solutions of (4.8). From Itô's formula, we can derive

$$\begin{aligned} E|y_t^1 - y_t^2|^2 + 2E \int_t^T \langle (A(s) + A_1(s))(y_s^1 - y_s^2), y_s^1 - y_s^2 \rangle + E \int_t^T |z_s^1 - z_s^2|^2 ds \\ = -2E \int_t^T (g(y_s^1, z_s^1, s) - g(y_s^2, z_s^2, s), y_s^1 - y_s^2) ds \\ - 2E \int_t^T \langle f(y_s^1, z_s^1, s) - f(y_s^2, z_s^2, s), y_s^1 - y_s^2 \rangle ds. \end{aligned}$$

Thus, by (2.8), (2.9), and (4.7), we obtain

$$\begin{aligned} E|y_t^1 - y_t^2|^2 + 2E \int_t^T \langle (A_1(s))(y_s^1 - y_s^2), y_s^1 - y_s^2 \rangle ds + 2\alpha E \int_t^T \|y_s^1 - y_s^2\|^2 ds \\ + E \int_t^T |z_s^1 - z_s^2|^2 ds \\ \leq 2\lambda E \int_t^T |y_s^1 - y_s^2|^2 ds + 2CE \int_t^T (|y_s^1 - y_s^2|^2 + |y_s^1 - y_s^2||z_s^1 - z_s^2|) ds \\ + 2E \int_t^T [C|y_s^1 - y_s^2| \|y_s^1 - y_s^2\| + |z_s^1 - z_s^2| \langle (2A_1(s))(y_s^1 - y_s^2), y_s^1 - y_s^2 \rangle^{1/2}] ds. \end{aligned}$$

Since $\|A_1(s)\|_{\mathcal{L}(V, V)} \leq C$, we can choose a small constant δ with $1 > 2\delta > 0$ such that

$$\frac{\delta}{1 - 2\delta} \langle 2A_1(s)y, y \rangle \leq \frac{\alpha}{2} \|y\|^2, \quad \forall y \in V.$$

Also, we have

$$2C|y^1 - y^2| \|y^1 - y^2\| < \frac{\alpha}{2} \|y^1 - y^2\|^2 + \frac{2}{\alpha} C^2 |y^1 - y^2|^2$$

and

$$\begin{aligned} & 2|z^1 - z^2| \langle 2A_1(s)(y^1 - y^2), y^1 - y^2 \rangle^{1/2} \\ & \leq (1 - 2\delta)|z^1 - z^2|^2 + \frac{1}{1 - 2\delta} \langle 2A_1(s)(y^1 - y^2), y^1 - y^2 \rangle, \\ & 2C|y^1 - y^2||z^1 - z^2| \leq C\delta^{-1}|y^1 - y^2|^2 + \delta|z^1 - z^2|^2. \end{aligned}$$

It follows that

$$\begin{aligned} & E|y_t^1 - y_t^2|^2 + \frac{\alpha}{2} E \int_t^T \|y_s^1 - y_s^2\|^2 ds + \delta E \int_t^T |z_s^1 - z_s^2|^2 ds \\ & \leq C_1 E \int_t^T |y_s^1 - y_s^2|^2 ds, \quad \forall t \in [0, T], \end{aligned}$$

with $C_1 = 2\lambda + 2C + 4\alpha^{-1}C^2 + \delta^{-1}C^2$. Thus we can apply Gronwall's inequality to obtain $y^1 = y^2, z^1 = z^2$.

(ii) *Existence.* Set $z_t^0 = 0$. According to Theorem 2.2, we can alternatively solve the following equation:

$$\begin{aligned} (4.9) \quad Y &= y_t^j + \int_t^T (A(s)y_s^j + A_1(s)y_s^j + f(y_s^j, z_s^{j-1}, s) + g(y_s^j, z_s^j, s)) ds \\ &+ \int_t^T z_s^j dW_s, \quad j = 1, 2, \dots \end{aligned}$$

By Itô's formula, we have

$$\begin{aligned} & E|y_t^{j+1} - y_t^j|^2 + 2E \int_t^T \langle (A(s) + A_1(s))(y_s^{j+1} - y_s^j), y_s^{j+1} - y_s^j \rangle \\ & \quad + E \int_t^T |z_s^{j+1} - z_s^j|^2 ds \\ &= -2E \int_t^T (g(y_s^{j+1}, z_s^{j+1}, s) - g(y_s^j, z_s^j, s), y_s^{j+1} - y_s^j) ds \\ & \quad - 2E \int_t^T \langle f(y_s^{j+1}, z_s^j, s) - f(y_s^j, z_s^{j-1}, s), y_s^{j+1} - y_s^j \rangle ds. \end{aligned}$$

Again, from (2.8), (2.9), and (4.7), we have

$$\begin{aligned} & E|y_t^{j+1} - y_t^j|^2 + 2E \int_t^T \langle A_1(s)(y_s^{j+1} - y_s^j), y_s^{j+1} - y_s^j \rangle ds \\ & \quad + 2\alpha E \int_t^T \|y_t^{j+1} - y_t^j\|^2 dt + E \int_t^T |z_s^{j+1} - z_s^j|^2 ds \\ & \leq CE \int_t^T (2\|y_t^{j+1} - y_t^j\| + |z_s^{j+1} - z_s^j|) |y_s^{j+1} - y_s^j| ds \\ & \quad + (2\lambda + 2C)E \int_t^T |y_s^{j+1} - y_s^j|^2 ds \\ & \quad + 2E \int_t^T |z_s^j - z_s^{j-1}| \langle 2A_1(s)(y_s^{j+1} - y_s^j), y_s^{j+1} - y_s^j \rangle^{1/2} ds. \end{aligned}$$

Taking δ as in the proof of the uniqueness and using the similar technique yield

$$(4.10) \quad \begin{aligned} & E|y_t^{j+1} - y_t^j|^2 + \frac{\alpha}{2} E \int_t^T \|y_s^{j+1} - y_s^j\|^2 dt + (1-\delta) E \int_t^T |z_s^{j+1} - z_s^j|^2 ds \\ & \leq C_2 E \int_t^T |y_s^{j+1} - y_s^j|^2 ds + (1-2\delta) E \int_t^T |z_s^j - z_s^{j-1}|^2 ds, \end{aligned}$$

where the constant C_2 is the same as in the proof of the uniqueness. Define

$$\begin{aligned} Y_s^j &= E \int_t^T |y_s^{j+1} - y_s^j|^2 ds, \quad j = 1, 2, \dots, \\ Z_s^j &= E \int_t^T |z_s^{j+1} - z_s^j|^2 ds, \quad j = 0, 1, 2, \dots. \end{aligned}$$

It follows from (4.10) that

$$(4.11) \quad -\frac{d}{dt} Y_t^j + (1-\delta) Z_t^j \leq C_2 Y_t^j + (1-2\delta) Z_t^{j-1} \quad Y_0^j = 0, \quad j = 1, 2, \dots,$$

or

$$-\frac{d}{dt} (Y_t^j e^{C_2 t}) + (1-\delta) e^{C_2 t} Z_t^j \leq (1-2\delta) e^{C_2 t} Z_t^{j-1} \quad Y_0^j = 0, \quad j = 1, 2, \dots.$$

Integrating from t to T yields

$$(4.12) \quad Y_t^j + (1-\delta) \int_t^T e^{C_2(s-t)} Z_s^j ds \leq (1-2\delta) \int_t^T e^{C_2(s-t)} Z_s^{j-1} ds.$$

It follows, in particular, that

$$\int_0^T e^{C_2 s} Z_s^j ds \leq \left(\frac{1-2\delta}{1-\delta} \right)^j K, \quad j = 1, 2, \dots,$$

with $K = E \int_0^T |z_s^1|^2 ds$; also then

$$(4.13) \quad Y_t^j \leq Y_0^j \leq \left(\frac{1-2\delta}{1-\delta} \right)^j K, \quad j = 1, 2, \dots.$$

From (4.11) and the fact that $dY_t^j dt \leq 0$, we derive that

$$Z_0^j \leq K_1 \left(\frac{1-2\delta}{1-\delta} \right)^j + \frac{(1-2\delta)}{(1-\delta)} Z_0^{j-1}, \quad j = 1, 2, \dots,$$

with $K_1 = C_2 K (1-\delta)^{-1}$. This implies

$$Z_0^j = E \int_0^T |z_s^{j+1} - z_s^j|^2 ds \leq \left(\frac{1-2\delta}{1-\delta} \right)^j (jK_1 + Z_0^0), \quad j = 1, 2, \dots.$$

It turns out that $\{(y^j, z^j)\}$ is a Cauchy sequence in $\mathcal{M}^2(H) \times \mathcal{M}^2(H)$. From (4.10), $\{y^j\}$ is also a Cauchy sequence in $\mathcal{M}^2(V)$ and $L^2(\Omega, C(0, T; H))$. Passing limit in (4.9) as $j \rightarrow \infty$, we obtain that the pair (y_t, z_t) defined by

$$\begin{aligned} y &= \lim_{j \rightarrow \infty} y^j \text{ in } \mathcal{M}^2(V), \\ z &= \lim_{j \rightarrow \infty} z^j \text{ in } \mathcal{M}^2(H) \end{aligned}$$

solves (4.8). \square

We now proceed to the proof of Theorem 4.1.

Proof of Theorem 4.1. Set

$$\begin{aligned}
 \langle A_1(t)\varphi_1, \varphi \rangle &= \frac{1}{2} \int_{R^n} (\sigma\sigma^*(x, t) D\varphi_1, D\varphi) dx, \quad \forall \varphi_1, \varphi \in V, \\
 \langle A(t)\varphi_1, \varphi \rangle &= \frac{1}{2} \int_{R^n} (\pi\pi^*(x, t) D\varphi_1, D\varphi) dx, \quad \forall \varphi_1, \varphi \in V, \\
 \langle f(\psi, t), \varphi \rangle &= \int_{R^n} (\psi(x), \sigma(x, t) D\varphi) dx, \quad \forall \psi \in H, \varphi \in V, \\
 g(\varphi, \psi, t) &= \frac{1}{2} \sum_{i,j=1}^n \partial_{x_i}(\sigma\sigma^*)_{ij} \partial_{x_j} \varphi \\
 &\quad + \frac{1}{2} \sum_{i,j=1}^n \partial_{x_i}(\pi\pi^*)_{ij} \partial_{x_j} \varphi - \psi \sum_{i=1}^n \partial_{x_i} \sigma_i(x, t) \\
 &\quad + \inf_{v \in U} \{(b(x, v, t), D\varphi(x)) + L(x, v, t)\}, \quad \forall \varphi \in V, \\
 \eta &= 0, \quad Y(\omega) = h(x, \omega).
 \end{aligned}$$

It is easy to see that

$$\begin{aligned}
 A_1(t), A_2(t) &: [0, T] \rightarrow \mathcal{L}(V; V'), \\
 f(\psi, t) &: H \times [0, T] \rightarrow V', \\
 g(\varphi, \psi, t) &: V \times H \times [0, T] \rightarrow H, \\
 Y(\omega) &: \Omega \rightarrow H.
 \end{aligned}$$

Thus, for each $w, \varphi \in V, \psi \in H$, we have that

$$\langle (A(s) + A_1(s))\varphi + f(\psi, s) + g(\varphi, \psi, s), w \rangle = \int_{R^n} H(D^2\varphi, D\varphi, D\psi, x, t) w(x) dx.$$

With the above notations, we can write the stochastic HJB equation (3.7) in the form of (4.8).

It remains to check that the conditions of Lemma 4.2 are satisfied. We will only verify (4.6) and (4.7) because the other conditions are easy to check. First, for (4.6), we need to verify

$$\begin{aligned}
 (4.14) \quad & \int_{R^n} \left| \inf_{v \in U} \{(b(x, v), D\varphi_1(x)) + L(x, v)\} \right. \\
 & \left. - \inf_{v \in U} \{(b(x, v), D\varphi_2(x)) + L(x, v)\} \right|^2 dx \leq C \|\varphi_2 - \varphi_1\|_V^2.
 \end{aligned}$$

For any $\varepsilon > 0$ and $\varphi_1, \varphi_2 \in V$, we can choose two measurable functions $v_l^\varepsilon(x): \mathbb{R}^n \rightarrow U$, $l = 1, 2$, such that

$$\begin{aligned}
 & \int_{R^n} |(b(x, v_l^\varepsilon, t), D\varphi_l(x)) + L(x, v_l^\varepsilon, t) - \inf_{v \in U} \{(b(x, v, t), D\varphi_l(x)) + L(x, v, t)\}|^2 \\
 & = \int_{R^n} |\delta_l^\varepsilon(x)|^2 dx \leq \varepsilon, \quad l = 1, 2.
 \end{aligned}$$

For each $x \in \mathbb{R}^n$, we have

$$\begin{aligned} & \inf_{v \in U} \{(b(x, v, t), D\varphi_1(x)) + L(x, v, t)\} - \inf_{v \in U} \{(b(x, v, t), D\varphi_2(x)) + L(x, v, t)\} \\ & \quad \cong (b(x, v_1^\varepsilon, t), D\varphi_1(x)) + L(x, v_1^\varepsilon, t) - (b(x, v_1^\varepsilon, t), D\varphi_2(x)) - L(x, v_1^\varepsilon, t) - |\delta_1^\varepsilon(x)| \\ & \quad \cong -C|D\varphi_1(x) - D\varphi_2(x)| - |\delta_1^\varepsilon(x)|. \end{aligned}$$

Similarly,

$$\begin{aligned} & \inf_{v \in U} \{(b(x, v, t), D\varphi_2(x)) + L(x, v, t)\} - \inf_{v \in U} \{(b(x, v, t), D\varphi_1(x)) + L(x, v, t)\} \\ & \quad \cong -C|D\varphi_1(x) - D\varphi_2(x)| - |\delta_2^\varepsilon(x)|. \end{aligned}$$

The above two inequalities imply

$$\begin{aligned} & \int_{\mathbb{R}^n} \left| \inf_{v \in U} \{(b(x, v, t), D\varphi_1(x)) + L(x, v, t)\} - \inf_{v \in U} \{(b(x, v, t), D\varphi_2(x)) + L(x, v, t)\} \right|^2 dx \\ & \quad \leq C \|\varphi_1 - \varphi_2\|_V^2 + |\delta_1^\varepsilon|_H^2 + |\delta_2^\varepsilon|_H^2 \\ & \quad \leq C \|\varphi_1 - \varphi_2\|_V^2 + 2\varepsilon. \end{aligned}$$

Let $\varepsilon \rightarrow 0$, we obtain (4.14).

Finally, the last Lipschitz condition in (2.9) can be derived as follows. For any $\psi_1, \psi_2 \in H$, $\varphi \in V$, $t \in [0, T]$, we have

$$\begin{aligned} |\langle f(\psi_1, t) - f(\psi_2, t), \varphi \rangle| &= \left| \int_{\mathbb{R}^n} (\psi_1 - \psi_2, \sigma^*(x, t) D\varphi) dx \right| \\ &\leq |\psi_1 - \psi_2|_H |\sigma^* D\varphi|_H \\ &= |\psi_1 - \psi_2|_H \left[\int_{\mathbb{R}^n} (\sigma \sigma^*(x, t) D\varphi, D\varphi) dx \right]^{1/2} \\ &= |\psi_1 - \psi_2|_H \langle A_1(t) \varphi, \varphi \rangle^{1/2}. \end{aligned} \quad \square$$

5. Linear quadratic case: Stochastic Riccati equation.

5.1. Problem formulation. As a particular but important case, we consider a linear stochastic control system with quadratic cost function and random coefficients. For simplicity, we assume that σ, π, b, L are all time-invariant. We assume also that W_t^1 is a one-dimensional Brownian motion ($p=1$). In this case all data in (3.1), (3.2) can be written as follows:

$$\begin{aligned} b(x, v) &= Ax + Bv, \\ \sigma(x, v) &= Cx + Dv, \\ \pi(x, v) &= Gx + Hv, \\ L(x, v) &= (Rx, x) + (Nv, v), \\ h(x, \omega) &= (Q(\omega)x, x), \end{aligned}$$

where A, C , and G are $n \times n$ matrices; B, D , and H are $n \times k$ matrices; $R \in S^n$; $N \in S^k$; $Q(\omega): \Omega \rightarrow S^n$. Here S^n (respectively, S^k) denotes the space of all $n \times n$ (respectively, $k \times k$ symmetric matrices), with scalar product $\langle Q_1, Q_2 \rangle = \text{tr}(Q_1, Q_2)$.

We assume

$$(5.1) \quad \begin{aligned} Q(\omega) &\text{ is } \mathcal{F}_T\text{-measurable, bounded, and nonnegative,} \\ R &\text{ is nonnegative, } N \text{ is positive.} \end{aligned}$$

The admissible controls now are valued in \mathbb{R}^k . Equation (2.1) now becomes

$$(5.2) \quad \begin{aligned} dy_s &= (Ay_s + Bv_s) ds + (Cy_s + Dv_s) dW_s + (Gy_s + Hv_s) dW_s^1, \\ 0 \leq s \leq t \leq T, \quad y_t &= x, \end{aligned}$$

and the value function is

$$(5.2) \quad \Phi_t(x) = \inf_{v \in \mathcal{M}^2(\mathbb{R}^k)} E^{\mathcal{F}_t} \left\{ \int_t^T [(Ry_s, y_s) + (Ny_s, y_s)] ds + (Qy_T, y_T) \right\}.$$

By classical methods, we can prove that $\Phi_t(x)$ is of the quadratic form

$$(5.3) \quad \Phi_t(x) = (K_t x, x), \quad \forall (x, t),$$

where K_t is an S^n -valued \mathcal{F}_t -adapted process. From § 3, we know that $\Phi_t(x)$ can be formally written as

$$(5.4) \quad \Phi_t(x) = (Qx, x) + \int_t^T \Gamma_s(x) ds - \int_t^T \Psi_s(x) dW_s,$$

where Γ_t and Ψ_t are \mathcal{F}_t -adapted real processes. Formally, we have

$$(5.5) \quad \Psi_t(x) = (M_t x, x),$$

where M_t is an \mathcal{F}_t -adapted process valued in S^n . If such (K_t, M_t) is bounded, we can use the dynamic programming principle to verify that the pair $(\Phi_t(x), \Psi_t(x))$ is a solution of the stochastic HJB equation (3.7).

We can formally derive the equation that characterizes (K_t, M_t) . Substituting (5.4), (5.5), and all linear or quadratic data b, σ, π, L, h into (3.7), from Itô's formula we can obtain

$$(5.6) \quad \begin{aligned} -dK_t &= [A^* K_t + K_t A + C^* K_t C + G^* K_t G + M_t C + C^* M_t \\ &\quad - \hat{B}(K_t, M_t) \hat{N}^{-1}(K_t) \hat{B}^*(K_t, M_t)] dt - M_t dW_t, \\ K_T &= Q, \end{aligned}$$

where for each $K, M \in S^n$, we denote

$$(5.7) \quad \begin{aligned} \hat{B} &= \hat{B}(K, M) = KB + MD + C^* KD + G^* KH, \\ \hat{N} &= \hat{N}(K) = N + D^* KD + H^* KH. \end{aligned}$$

Since $\hat{B} \hat{N}^{-1} \hat{B}$ are nonlinear with respect to K and M , (5.6) is a nonlinear backward stochastic equation. Equation (5.6) is called a stochastic matrix Riccati equation. When Q is deterministic, it becomes an ordinary (deterministic) Riccati equation.

5.2. Existence and uniqueness result. We can regard (5.6) as a nonlinear backward stochastic equation in the form (2.2), defined in the Euclidean space S^n . Theorem 2.1 cannot be applied directly, however, because $\hat{B} \hat{N}^{-1} \hat{B}$ does not satisfy the global Lipschitz condition. This kind of equation was first investigated by Bismut [4]. He obtained an existence theorem for the case where $C = 0, D = 0$ by a method based on the fixed point theorem. We will treat the case where only $D = 0$. In this case the

bounded variation part of K_t contains M_t . We must use a different method to overcome this difficulty. The case where D is nonzero is still an open problem.

When $D=0$, \hat{B} and \hat{N} become

$$(5.8) \quad \begin{aligned} \hat{B}(K, t) &= KB + G^*KH, \\ \hat{N}(K, t) &= N + H^*KH. \end{aligned}$$

We assert the following.

THEOREM 5.1. *We assume (5.1) and $D=0$. Then there exists a pair (K_t, M_t) in $(\mathcal{M}^2(S^n))^2$ satisfying the stochastic Riccati equation (5.6) such that K_t is nonnegative and bounded.*

To prove this theorem, the following lemma is in order.

LEMMA 5.2. *Let $\hat{A}_t, \hat{C}_t, \hat{G}_t$ be $\mathbb{R}^{n \times n}$ -valued, and \hat{R}_t be S^n -valued, \mathcal{F}_t -adapted processes. Assume that they are all bounded. Let \hat{Q} be a bounded \mathcal{F}_T -measurable random variable with values in S^n . Then there exists a pair (\hat{K}_t, \hat{M}_t) in $(\mathcal{M}^2(S^n))^2$ satisfying the following linear equation:*

$$(5.9) \quad \begin{aligned} -d\hat{K}_t &= [\hat{A}_t^* \hat{K}_t + \hat{K}_t \hat{A}_t + C_t^* \hat{K}_t C_t + \hat{G}_t^* \hat{K}_t \hat{G}_t \\ &\quad + (\hat{M}_t \hat{C}_t + \hat{C}_t^* \hat{M}_t) + \hat{R}_t], \end{aligned}$$

$$dt - \hat{M}_t dW_t, \quad \hat{K}_T = \hat{Q}.$$

Moreover,

$$(5.10) \quad \sup_{t, \omega} |\hat{K}_t(\omega)|^2 \leq k_0,$$

where the constant k_0 only depends on

$$\sup_{t, \omega} (|\hat{A}_t| + |\hat{C}_t| + |\hat{G}_t|),$$

and

$$\sup_{\omega} \left(|\hat{Q}|^2 + \int_0^T |\hat{R}_t|^2 \right) (\omega).$$

If \hat{R}_t, \hat{Q} are nonnegative, almost surely, then \hat{K}_t is also nonnegative, almost surely.

Proof. The above SDE can be regarded as an ordinary backward SDE in the Euclidean space S^n . According to Theorem 2.1, the existence and uniqueness as well as (5.10) hold. It remains to prove the nonnegativity of \hat{K} . For given (x, t) let y_s be the solution of

$$\begin{aligned} dy_s &= \hat{A}_s y_s ds + \hat{C}_s y_s dW_s + \hat{G}_s y_s dW_s^1, \\ y_t &= x, \quad t \leq s \leq T. \end{aligned}$$

Then, we can apply Itô's formula

$$d(\hat{K}_s y_s, y_s) = -(\hat{R}_s y_s, y_s) ds + (\hat{M}_s y_s, y_s) dW_s + 2(\hat{K}_s y_s, \hat{C}_s y_s dW_s + \hat{G}_s y_s dW_s^1).$$

Thus

$$(\hat{K}_t x, x) = E^{\mathcal{F}_t} \left[\int_t^T (\hat{R}_s y_s, y_s) ds + (\hat{Q} y_T, y_T) \right].$$

It follows that K_t is nonnegative whenever \hat{R}_s, \hat{Q}_s are nonnegative. \square

We now proceed to prove Theorem 5.1.

Proof of Theorem 5.1. (i) *Existence.* We define $F(K, M, U): S^n \times S^n \times \mathcal{L}(\mathbb{R}^n; \mathbb{R}^k) \rightarrow S^n$ by

$$\begin{aligned} F(K, M, U) &= (A + BU)^* K + K(A + BU) + C^* K C \\ &\quad + (G + HU)^* K (G + HU) + (MC + C^* M^*). \end{aligned}$$

We also define $\hat{U}(K): (S^n)^+ \rightarrow \mathcal{L}(\mathbb{R}^n; \mathbb{R}^k)$ by

$$\hat{U}(K) = -\hat{N}^{-1}(K)\hat{B}(K),$$

where $(S^n)^+$ is the set of nonnegative elements of S^n . With these notations, we can rewrite (5.6) as

$$(5.11) \quad \begin{aligned} -dK_t &= [F(K_t, M_t, \hat{U}(K_t)) + \hat{U}^*(K_t)N\hat{U}(K_t)] dt - M_t dW_t, \\ K_T &= Q. \end{aligned}$$

It is seen that

$$(5.12) \quad F(K, M, \hat{U}(K)) + \hat{U}^*(K)N\hat{U}(K) \leq F(K, M, U) + U^*NU, \quad \forall U \in \mathcal{L}(\mathbb{R}^n; \mathbb{R}^k).$$

We now iteratively construct a sequence of approximating solutions. First, we define (K_1, M_1) by solving the following linear backward SDE:

$$\begin{aligned} -dK_{1,t} &= F(K_{1,t}, M_{1,t}, 0) dt - M_{1,t} dW_t, \\ K_{1,T} &= Q. \end{aligned}$$

By Lemma 5.2, we can easily check that the above solution exists, and that $K_{1,t}$ is bounded and nonnegative. Thus $\hat{U}(K_{1,t})$ has meaning and is bounded. Then, we define (K_2, M_2) by solving

$$\begin{aligned} -dK_{2,t} &= [F(K_{2,t}, M_{2,t}, \hat{U}(K_{1,t})) + \hat{U}^*(K_{1,t})N\hat{U}(K_{1,t})] dt - M_{2,t} dW_t, \\ K_{2,T} &= Q. \end{aligned}$$

Again from Lemma 5.2, there is a unique solution (K_2, M_2) , which is bounded and nonnegative. Thus $\hat{U}(K_{2,t})$ is well defined and bounded. Inductively, we can define (K_{j+1}, M_{j+1}) , which is the unique bounded and nonnegative solution of

$$(5.13) \quad \begin{aligned} -dK_{j+1,t} &= F(K_{j+1,t}, M_{j+1,t}, \hat{U}(K_{j,t})) dt + \hat{U}^*(K_{j,t})N\hat{U}(K_{j,t}) dt - M_{j+1,t} dW_t, \\ K_{j+1,T} &= Q, \quad j = 1, 2, \dots \end{aligned}$$

We claim that the sequence $\{K_{j,t}\}$ is nonincreasing. Indeed, we have

$$\begin{aligned} -d(K_{j,t} - K_{j+1,t}) &= [F(K_{j,t}, M_{j,t}, \hat{U}(K_{j,t})) - F(K_{j+1,t}, M_{j+1,t}, \hat{U}(K_{j,t}))] dt \\ &\quad + [F(K_{j,t}, M_{j,t}, \hat{U}(K_{j-1,t})) + \hat{U}^*(K_{j-1,t})N\hat{U}(K_{j-1,t}) \\ &\quad - F(K_{j,t}, M_{j,t}, \hat{U}(K_{j,t})) - \hat{U}^*(K_{j,t})N\hat{U}(K_{j,t})] dt \\ &\quad - (M_{j,t} - M_{j+1,t}) dW_t \\ &= F(K_{j,t} - K_{j+1,t}, M_{j,t} - M_{j+1,t}, \hat{U}(K_{j,t})) dt \\ &\quad + R_{j,t} dt - (M_{j,t} - M_{j+1,t}) dW_t, \\ K_{j,T} - K_{j+1,T} &= 0, \end{aligned}$$

with

$$\begin{aligned} R_{j,t} &= F(K_{j,t}, M_{j,t}, \hat{U}(K_{j-1,t})) + \hat{U}^*(K_{j-1,t})N\hat{U}(K_{j-1,t}) \\ &\quad - F(K_{j,t}, M_{j,t}, \hat{U}(K_{j,t})) - \hat{U}^*(K_{j,t})N\hat{U}(K_{j,t}). \end{aligned}$$

From (5.12), R_j is nonnegative. Thus, according to Lemma 5.2, $K_{j,t} - K_{j+1,t}$ is also nonnegative. This implies that $\{K_{j,t}\}$ is the nonincreasing sequence

$$CI \geq K_{1,t} \geq K_{2,t} \geq \dots \geq K_{j,t} \geq \dots \geq 0.$$

It follows that $\{K_{j,t}\}$ converges almost surely to a nonnegative, S^n -valued process K_t . According to Lebesgue's convergence theorem, we have

$$E \int_0^T |K_{j,t} - K_t|^q \rightarrow 0, \quad \text{as } j \rightarrow \infty, \quad \forall q > 0.$$

Thus $\{K_{j,t}\}$, and also then $\{U(K_{j,t})\}$ is a Cauchy sequence in the above sense. We have also almost everywhere

$$E|K_{j,t} - K_t|^q \rightarrow 0, \quad \text{as } j \rightarrow \infty, \quad \forall q > 0.$$

By definition (5.13), we can apply Itô's formula to $|K_{k,t} - K_{j,t}|^2$,

$$\begin{aligned} & E|K_{k,0} - K_{j,0}|^2 + E \int_0^T |M_{k,t} - M_{j,t}|^2 dt \\ &= 2E \int_0^T \text{tr}[(K_{k,t} - K_{j,t})((M_{k,t} - M_{j,t})C + C^*(M_{k,t} - M_{j,t}))] dt + R(j, k), \\ &\leq \frac{1}{2} E \int_0^T |M_{k,t} - M_{j,t}|^2 dt + CE \int_0^T |K_{k,t} - K_{j,t}|^2 dt + R(j, k), \end{aligned}$$

where $R(j, k) \rightarrow 0$ as $\min(j, k) \rightarrow \infty$. Thus $\{M_{j,t}\}$ is a Cauchy sequence in $\mathcal{M}^2(\mathcal{L}(\mathbb{R}^d; S^n))$. Passing to the limit in (5.13), we obtain that (K_t, M_t) is a solution of (5.6), with

$$M_t = \lim_{j \rightarrow \infty} M_{j,t} \quad \text{in } \mathcal{M}^2(S^n).$$

(ii) *Uniqueness.* Let (K_t, M_t) and (K'_t, M'_t) be two pairs in $\mathcal{M}^2(S^n) \times \mathcal{M}^2(S^n)$ satisfying (5.6) (or (5.11)), such that K_t, K'_t are nonnegative and bounded. Then $\hat{U}(K_t)$ and $\hat{U}(K'_t)$ are well defined and bounded. We have

$$\begin{aligned} -d(K_t - K'_t) &= F(K_t - K'_t, M_t - M'_t, \hat{U}(K_t)) dt \\ &\quad + [F(K'_t, M'_t, \hat{U}(K_t)) + \hat{U}^*(K_t)N\hat{U}(K_t) \\ &\quad - F(K'_t, M'_t, \hat{U}(K'_t)) + \hat{U}^*(K'_t)N\hat{U}(K'_t)] dt - M_t dW_t, \\ K_T - K'_T &= 0, \end{aligned}$$

or

$$\begin{aligned} -d(K_t - K'_t) &= F(K_t - K'_t, M_t - M'_t, \hat{U}(K_t)) dt \\ &\quad + R' dt + (M_t - M'_t) dW_t, \\ K_T - K'_T &= 0, \end{aligned}$$

with

$$R' = F(K'_t, M'_t, \hat{U}(K_t)) + \hat{U}^*(K_t)N\hat{U}(K_t) - F(K'_t, M'_t, \hat{U}(K'_t)) + \hat{U}^*(K'_t)N\hat{U}(K'_t).$$

By (5.12), R' is nonnegative. It follows from Lemma 5.2 that $K_t - K'_t$ is nonnegative. Similarly, we can obtain that $K'_t - K_t$ is nonnegative. This implies $K_t = K'_t$. Consequently (from the uniqueness part of Lemma 5.2), $M_t = M'_t$. \square

As we mentioned in § 2 (verification theorem), once we obtain the solution of (5.6), which is regular enough, then the value function can be obtained by $\Phi_t = (K_t x, x)$. Moreover, the optimal feedback control can also be given. Specifically, we have the following corollary.

COROLLARY 5.3. *Let the assumptions of Theorem 5.1 hold. Then the value function is equal to $(K_t x, x)$. The optimal feedback control is*

$$u_t(x) = \hat{U}(K_t)x.$$

Proof. Since K_t is bounded and positive, $\hat{U}(K_t)$ is well posed and bounded. Let (x_0, t_0) be any given initial data. We want to minimize

$$J_{x_0, t_0}(v) = E \int_{t_0}^T [(Rx_t, x_t) + (N_t v_t, v_t)] ds + E(Qx_T, x_T)$$

subject to

$$dx_t = (Ax_t + Bv_t) dt + Cx_t dW_t + (Gx_t + Hv_t) dW_t^1, \quad x_{t_0} = x_0.$$

For any given admissible control v_t , we can apply Itô's formula to $(K_t x_t, x_t)$, and thus verify that

$$(K_{t_0} x_{t_0}, x_{t_0}) \leq E^{\mathcal{F}_{t_0}} \left\{ \left[(Qx_T, x_T) + \int_{t_0}^T (Rx_t, x_t) + (N_t v_t, v_t) \right] ds \right\}.$$

On the other hand, let y_t be a solution of

$$dy_t = (A + B\hat{U}(K_t))y_t dt + Cy_t dW_t + (G + H\hat{U}(K_t))y_t dW_t^1, \quad y_{t_0} = x_0.$$

We can again apply Itô's formula to $(K_t y_t, y_t)$,

$$(K_{t_0} x_{t_0}, x_{t_0}) = E^{\mathcal{F}_{t_0}} \left\{ \left[(Qy_T, y_T) + \int_{t_0}^T (Ry_t, y_t) + (N_t \hat{U}(K_t)y_t, \hat{U}(K_t)y_t) \right] ds \right\}.$$

It follows that, for almost surely (x_0, t_0) $(K_{t_0} x_0, x_0)$ is equal to the value function. Furthermore, if we set $u_t = \hat{U}(K_t)y_t$,

$$\begin{aligned} J_{x_0, t_0}(v) &= E \int_{t_0}^T [(Rx_t, x_t) + (N_t v_t, v_t)] ds + E(Qx_T, x_T) \\ &= EE^{\mathcal{F}_{t_0}} \left[\int_{t_0}^T [(Rx_t, x_t) + (N_t v_t, v_t)] ds + (Qx_T, x_T) \right] \\ &\geq EE^{\mathcal{F}_{t_0}} \left[\int_{t_0}^T [(Ry_t, y_t) + (N_t u_t, u_t)] ds + (Qy_T, y_T) \right] \\ &= J_{x_0, t_0}(u) \end{aligned}$$

It follows that u_t is the optimal control and so $U(K_t)x$ is the optimal feedback. \square

Acknowledgments. The author thanks the referees and the editors for their helpful suggestions, which made the revised version of this paper more readable.

REFERENCES

- [1] A. BENSOUSSAN, *Lectures on stochastic control*, in Nonlinear Filtering and Stochastic Control, Proceedings, Cortona, 1981, Lecture Notes in Math., 972, Springer-Verlag, New York, Berlin, pp. 1-62.
- [2] ———, *Stochastic maximum principle for distributed parameter system*, J. Franklin Inst., 315 (1983), pp. 387-406.
- [3] A. BENSOUSSAN AND J. L. LIONS, *Applications des Equations Variationnelles en Contrôle Stochastique*, Dunod, Paris, 1973.

- [4] J. M. BISMUT, *Linear quadratic optimal stochastic control with random coefficients*, SIAM J. Control Optim., 14 (1976), pp. 419-444.
- [5] ———, *An introductory approach to duality in optimal stochastic control*, SIAM Rev., 20 (1978), pp. 62-78.
- [6] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, Berlin, New York, 1975.
- [7] U. G. HAUSSMANN, *General necessary conditions for optimal control of stochastic systems*, Math. Programming Study, 6 (1976), pp. 34-48.
- [8] Y. HU AND S. PENG, *Adapted solution of backward stochastic evolution equation*, Stochastic Anal. Appl., to appear.
- [9] ———, *Maximum principle for semilinear stochastic evolution systems*, Stochastics, to appear.
- [10] H. KUNITA, *Ecole d'Eté de Probabilité de Saint-Flour XII*, Springer-Verlag, Berlin, New York, 1982.
- [11] N. V. KRYLOV, *Controlled Diffusion Processes*, Springer-Verlag, Berlin, New York, 1980.
- [12] ———, *Nonlinear Elliptic and Parabolic Equations of the Second Order*, D. Reidel, Dordrecht, 1987.
- [13] H. J. KUSHNER, *Necessary conditions for continuous parameter stochastic optimization problems*, SIAM J. Control., 10 (1972), pp. 550-565.
- [14] P. L. LIONS, *Optimal Control of Diffusion Processes and Hamilton Jacobi Equations, Part II*, Comm. Partial Differential Equations, 8 (1983), pp. 1229-1276.
- [15] E. PARDOUX, *Equations aux dérivées partielles stochastiques non linéaires monotones*, Thèse d'Etat à l'Université Paris Sud, Paris, FR, 1975.
- [16] E. PARDOUX AND S. PENG, *Adapted solution of backward stochastic equation*, Systems Control Lett., 14 (1990), pp. 55-61.
- [17] S. PENG, *A general stochastic maximum principle for optimal control problems*, SIAM J. Control Optim., 28 (1990), pp. 966-979.
- [18] ———, *A kind of HJB equations with random coefficients*, in Proceedings Nat. Conf. Control Theory, Xian, China, 1989, pp. 333-336. (In Chinese.)

A PENALIZATION METHOD FOR OPTIMAL CONTROL OF ELLIPTIC PROBLEMS WITH STATE CONSTRAINTS*

MAÏTINE BERGOUNIOUX†

Abstract. In this paper boundary or distributed stationary control problems are studied in relation to an elliptic operator and state and control constraints. Different kinds of conditions are formulated to prove the existence of a decoupled optimality system and Lagrange multipliers.

Key words. optimal control, optimality conditions, state constraints, Lagrange multipliers, elliptic operators

AMS(MOS) subject classification. 49B22

1. Introduction. In this paper we study state-constrained control problems for elliptic systems. Usually, the state constraints are pointwise constraints, but we do not need to specify them. We prove (for “classical” cases) the existence of a decoupled optimality system and Lagrange multipliers. The method and results are presented by simple examples of distributed or boundary control. It may be applied to many other systems (we must then choose a suitable functional frame). This question has been studied by many authors and, especially for distributed control systems, by Bonnans and Casas [7]–[10], [12], [13]. Their results concern the choice of a “good” functional space. Assuming the Slater qualification constraint, they obtain a first-order optimality system. Lasiecka [16], Barbu [4], and Mackenroth [20], [21] have studied parabolic problems (with a dual approach), White [23] hyperbolic problems, Abergel [1] ill-posed problems, and, with Temam [2], nonqualified problems.

Most of these methods consist in the “relaxation” of the convex (i.e., the state) constraints and use classical results of convex analysis. More precisely, let X, Y be two Banach spaces, $T: X \rightarrow Y$ a continuous linear map, $h: X \rightarrow \bar{\mathbb{R}}$, and $g: Y \rightarrow \bar{\mathbb{R}}$ lower semicontinuous (l.s.c.) convex function. We know [14], [22] that a solution (if it exists) \bar{x} of

$$(1.1) \quad \min h(x) + g(Tx), \quad x \in X,$$

also satisfies

$$(1.2) \quad \partial h(\bar{x}) + T^* \partial g(T\bar{x}) \ni 0$$

(that is, the first-order optimality system) if the following qualification condition is ensured:

$$(1.3) \quad \exists y \text{ in the range of } T; \quad g \text{ is continuous at } y.$$

In the cases we consider, (1.3) is realized as soon as the following condition is satisfied:

$$(1.4) \quad \text{Int}_Y [T(\text{dom}(f)) - \text{dom}(g)] \ni 0.$$

We do not use such methods; we keep the state constraints and penalize the relation linking the state y to the control v via the partial differential equations (and, if necessary, via the boundary condition). The method is classical (Lions [17]) and is generally used for the study of singular systems (especially for multistate systems [18]). We prove the existence of a decoupled optimality system, assuming, of course, a

* Received by the editors November 22, 1989; accepted for publication (in revised form) January 14, 1991.

† Département de Mathématiques, Université d'Orléans, B.P. 6759, 45046 Orléans Cédex 2, France.

qualification hypothesis. We retrieve already-known results and then try to weaken them by changing the topology of the test function spaces (see [6]).

2. Formulation of the problem. Let Ω be an open bounded subset of \mathbb{R}^n with C^∞ boundary Γ . Let us consider the following system:

$$(2.1) \quad \mathbf{P}(u, v) \begin{cases} Ay = u & \text{in } \Omega, \\ By = v & \text{on } \Gamma, \end{cases}$$

where

- (i) $u \in \mathbb{L}^2(\Omega)$, $v \in \mathbb{L}^2(\Gamma)$;
- (ii) A is a differential elliptic operator defined by

$$(2.2) \quad \begin{aligned} Ay &= - \sum_{i,j=1}^n \partial_{x_i} (a_{ij}(x) \partial_{x_j} y) + a_o(x)y \quad \text{with} \\ a_{ij}, a_o &\in \mathcal{C}^2(\bar{\Omega}) \text{ for } i, j = 1, \dots, n, \quad \inf_{x \in \bar{\Omega}} a_o(x) > 0, \end{aligned}$$

$$\sum_{ij=1}^n a_{ij}(x) \xi_i \xi_j \geq c \sum_{i=1}^n \xi_i^2 \quad \forall x \in \bar{\Omega}, \quad \forall \xi \in \mathbb{R}^n, c > 0;$$

- (iii) B is a boundary operator

$$(2.3) \quad \begin{aligned} By &= y|_{\Gamma} && \text{Dirichlet boundary conditions,} \\ By &= \frac{\partial y}{\partial \nu_A} && \text{Neumann boundary conditions,} \end{aligned}$$

where $\partial y / (\partial \nu_A)$ is the usual normal derivative (associated with A).

More precisely, we study the following two cases:

1. The control is *distributed*, so that $\mathbf{P}(u, v)$ becomes

$$\mathbf{P}(v) \begin{cases} Ay = f + v & \text{in } \Omega, \\ By = 0 & \text{on } \Gamma; \end{cases}$$

in this case, the control space X is $\mathbb{L}^2(\Omega)$ ($f \in \mathbb{L}^2(\Omega)$);

2. The control is a *boundary* one, so that $\mathbf{P}(u, v)$ becomes

$$\mathbf{P}(v) \begin{cases} Ay = f & \text{in } \Omega, \\ By = v & \text{on } \Gamma; \end{cases}$$

in this case, the control space X is $\mathbb{L}^2(\Gamma)$ ($f \in \mathbb{L}^2(\Omega)$).

We call T the affine application from X to $\mathbb{L}^2(\Omega)$ such that $y = T(v)$ is the unique solution of $\mathbf{P}(v)$ (in all the cases we consider, $\mathbf{P}(v)$ has a unique solution).

Let $J: \mathbb{L}^2(\Omega) \times X \rightarrow \mathbb{R}^+$ be a strictly convex, Gâteaux-differentiable functional (cost functional). We may choose, for example,

$$(2.4) \quad J(y, v) = \frac{1}{2} \|y - z_d\|_{\Omega}^2 + \frac{1}{2} \|v\|_X^2,$$

where $z_d \in \mathbb{L}^2(\Omega)$, N is a nonnegative real, $\|\cdot\|_X$ is the X -norm, and $\|\cdot\|_{\Omega}$ is the $\mathbb{L}^2(\Omega)$ -norm.

Let K be a nonempty closed convex subset of $\mathbb{L}^2(\Omega)$, and let U_{ad} be a nonempty closed convex subset of X .

Remark 1. K may be a priori any convex set, but we are essentially interested in *pointwise constraints*. So we must be able to define $y(x)$ everywhere on Ω . Let us consider, for example, the following distributed case: if the dimension n is less than

or equal to 3, it is (generally) sufficient to choose the control (and f) in $\mathbb{L}^2(\Omega)$ because most of the time the state y belongs to $\mathbb{H}^2(\Omega) \subseteq \mathcal{C}(\bar{\Omega})$ (due to Sobolev embeddings). For any dimension n , we should take \mathbb{L}^p as the control space, with $p > n/2$, so that $y(\in W^{2,p})$ still belongs to $\mathcal{C}(\bar{\Omega})$ (see [3]). Then a suitable choice of J may be, for example, $J(y, v) = \frac{1}{2}\|y - z_d\|_{\Omega}^2 + N/p\|v\|_X^p$. Anyway, the method can be developed in the same way without any difficulties.

We consider the following optimal control problem:

$$\begin{aligned} (\Pi) \quad & \text{Min } J(y, v), \\ & y = T(v), \\ & y \in K, v \in U_{\text{ad}}. \end{aligned}$$

Remark 2. When $f=0$, this problem may be considered as a particular case of (1.1) with X as the control space, $Y \subseteq \mathbb{L}^2(\Omega)$, $h = J \circ \mathcal{T}$, and $g = \mathbf{1}_K$. ($\mathcal{T}(v) = (T(v), v)$ for each v in X and $\mathbf{1}_K$ is the indicatrix of the state constraints set.) Then (1.4) is implied by the classical Slater condition [7], [13]

$$(\mathcal{S}) \quad \exists v_o \in U_{\text{ad}} \text{ such that } T(v_o) \in \text{Int}_Y(K).$$

We must suppose that the feasible domain \mathcal{D} of (II) is nonempty, i.e.,

$$(2.5) \quad \exists v_o \in U_{\text{ad}} \text{ such that } y_o = T(v_o) \in K$$

$$(\mathcal{D} = \{(y, v) \in K \times U_{\text{ad}} \mid y = T(v)\}).$$

$K \times U_{\text{ad}}$ is convex, closed, and the application from X to $\mathbb{L}^2(\Omega)$, $v \mapsto T(v)$ is affine. So \mathcal{D} is nonempty, closed, and convex; moreover, J is strictly convex, so (II) has a unique solution (\bar{y}, \bar{v}) .

Let us differentiate J as follows:

$$\forall v \in U_{\text{ad}} \text{ such that } T(v) \in K: J'(\bar{y}, \bar{v})(T(v) - T(\bar{v}), v - \bar{v}) \geq 0.$$

For example, when J is given by (2.4), and $X = \mathbb{L}^2(\Omega)$, this optimality condition can be written as follows:

$$(2.6) \quad \begin{aligned} & \forall v \in U_{\text{ad}}, \\ & \forall y \in K \text{ s.t. } y = T(v) \quad \int_{\Omega} (\bar{y} - z_d)(y - \bar{y}) \, dx + N \int_{\Omega} \bar{v}(v - \bar{v}) \, dx \geq 0. \end{aligned}$$

Thus it is easy to obtain a coupled optimality condition (even with state constraints). “Coupled” means that the test functions y and v are linked with the relation $y = T(v)$. The optimal solution of (II) is obtained by projection on the convex \mathcal{D} , quite difficult to “describe” (especially for numerical tests). The problem now is to obtain a decoupled system where the test functions v and y may be chosen in U_{ad} and K , independently of each other.

3. Case I. Distributed control.

3.1. Dirichlet boundary conditions.

3.1.1. Penalization of the problem (II). Let us consider the following *Dirichlet* problem:

$$(3.1) \quad \mathbf{P}(v) \begin{cases} Ay = f + v & \text{in } \Omega, \\ y = 0 & \text{on } \Gamma, \end{cases}$$

where $v \in \mathbb{L}^2(\Omega)$ and $f \in \mathbb{L}^2(\Omega)$. It is well known that $\mathbf{P}(v)$ has a unique solution $y(v)$

in $\mathbb{H}^2(\Omega) \cap \mathbb{H}_o^1(\Omega)$ (cf. [17], [18]) and T is affine and continuous from $\mathbb{L}^2(\Omega)$ to $\mathbb{H}^2(\Omega) \cap \mathbb{H}_o^1(\Omega)$. (If $f=0$, T is an isomorphism.)

In this case,

$$J(y, v) = \frac{1}{2} \int_{\Omega} (y - z_d)^2 dx + \frac{1}{2} \int_{\Omega} v^2 dx,$$

and K (respectively, U_{ad}) is a nonempty, closed, convex subset of $\mathbb{H}^2(\Omega) \cap \mathbb{H}_o^1(\Omega)$ (respectively, $\mathbb{L}^2(\Omega)$).

We are going to penalize (Π) (cf. [18]). Let us choose $\varepsilon > 0$ and define J_ε on $[\mathbb{H}^2(\Omega) \cap \mathbb{H}_o^1(\Omega)] \times \mathbb{L}^2(\Omega)$ by

$$(3.2) \quad J_\varepsilon(y, v) = J(y, v) + \frac{1}{2\varepsilon} \int_{\Omega} (Ay - v - f)^2 dx.$$

Let (Π_ε) be the penalized problem

$$(\Pi_\varepsilon) \quad \begin{aligned} &\text{Min } J_\varepsilon(y, v), \\ &y \in K, v \in U_{ad}. \end{aligned}$$

PROPOSITION 3.1. *Let $\varepsilon > 0$. There is a unique couple $(y_\varepsilon, v_\varepsilon)$ in $K \times U_{ad}$ optimal solution of (Π_ε) and for all $(y, v) \in K \times U_{ad}$, $J'_\varepsilon(y_\varepsilon, v_\varepsilon)(y - y_\varepsilon, v - v_\varepsilon) \geq 0$.*

Proof. J_ε is a coercive, strictly convex functional minimized on a nonempty closed, convex domain, so we get the existence of $(y_\varepsilon, v_\varepsilon)$. Moreover, J is differentiable, and we have the wanted inequality. \square

This means that

$$\begin{aligned} &\int_{\Omega} (y_\varepsilon - z_d)(y - y_\varepsilon) dx + N \int_{\Omega} v_\varepsilon(v - v_\varepsilon) dx + \frac{1}{\varepsilon} \int_{\Omega} (Ay_\varepsilon - v_\varepsilon - f)[A(y - y_\varepsilon)] dx \\ &\quad - \frac{1}{\varepsilon} \int_{\Omega} (Ay_\varepsilon - v_\varepsilon - f)(v - v_\varepsilon) dx \geq 0. \end{aligned}$$

As y and v are independent variables, this optimality condition can be decoupled to obtain the following system:

$$(3.3) \quad \forall y \in K \quad \int_{\Omega} (y_\varepsilon - z_d)(y - y_\varepsilon) dx + \frac{1}{\varepsilon} \int_{\Omega} [Ay_\varepsilon - v_\varepsilon - f][A(y - y_\varepsilon)] dx \geq 0,$$

$$(3.4) \quad \forall v \in U_{ad} \quad \int_{\Omega} \left[Nv_\varepsilon - \frac{1}{\varepsilon} (Ay_\varepsilon - v_\varepsilon - f) \right] (v - v_\varepsilon) dx \geq 0.$$

Let us call

$$(3.5) \quad q_\varepsilon = \frac{Ay_\varepsilon - v_\varepsilon - f}{\varepsilon} \in \mathbb{L}^2(\Omega),$$

and let $p_\varepsilon \in \mathbb{H}_o^1(\Omega) \cap \mathbb{H}^2(\Omega)$ be the adjoint state solution of

$$(3.6) \quad \begin{aligned} A^*p_\varepsilon &= y_\varepsilon - z_d && \text{in } \Omega, \\ p_\varepsilon &= 0 && \text{on } \Gamma, \end{aligned}$$

where A^* is the adjoint of A .

Remark 3. If A is given by (2.2), A^* is an isomorphism from $\mathbb{H}^2(\Omega) \cap \mathbb{H}_o^1(\Omega)$ onto $\mathbb{L}^2(\Omega)$.

Let us recall Green's formula (cf. [18]). For all $(y, z) \in \mathbb{H}^2(\Omega) \times \mathbb{H}^2(\Omega)$,

$$(3.7) \quad \int_{\Omega} (Ay)z \, dx = \int_{\Omega} y(A^*z) \, dx - \int_{\Gamma} \frac{\partial y}{\partial \nu_A} z \, d\Gamma + \int_{\Gamma} \frac{\partial z}{\partial \nu_{A^*}} y \, d\Gamma.$$

As (3.3) is equivalent to

$$\int_{\Omega} (A^*p_{\varepsilon})(y - y_{\varepsilon}) \, dx + \int_{\Omega} q_{\varepsilon}[A(y - y_{\varepsilon})] \, dx \geq 0, \quad \forall y \in K,$$

we obtain the following result.

PROPOSITION 3.2. *Let $(y_{\varepsilon}, v_{\varepsilon})$ be the solution of (Π_{ε}) ,*

$$(3.8) \quad \forall y \in K \quad \int_{\Omega} (q_{\varepsilon} + p_{\varepsilon})[A(y - y_{\varepsilon})] \, dx \geq 0,$$

$$(3.9) \quad \forall v \in U_{\text{ad}} \quad \int_{\Omega} (Nv_{\varepsilon} - q_{\varepsilon})(v - v_{\varepsilon}) \, dx \geq 0.$$

Now we must study the asymptotic behaviour of these relations. First, we can describe the behaviour of y_{ε} , v_{ε} , and p_{ε} as ε tends to zero.

THEOREM 3.3. *Let (\bar{y}, \bar{v}) be the solution of (2). v_{ε} converges to \bar{v} strongly in $\mathbb{L}^2(\Omega)$, and y_{ε} converges to \bar{y} strongly in $\mathbb{H}^2(\Omega) \cap \mathbb{H}_o^1(\Omega)$, as ε tends to zero.*

Proof. The proof is nearly the same as the one given by Lions [17] for a similar penalization problem.

Let us give $\varepsilon > 0$. We have

$$0 \leq J_{\varepsilon}(y_{\varepsilon}, v_{\varepsilon}) \leq J_{\varepsilon}(\bar{y}, \bar{v}) = J(\bar{y}, \bar{v}) = j < +\infty.$$

Therefore there exists $k_o \geq 0$ such that

$$\sup_{\varepsilon > 0} \|y_{\varepsilon}\|_{\Omega} \leq k_o \quad \text{and} \quad \sup_{\varepsilon > 0} \|v_{\varepsilon}\|_{\Omega} \leq k_o.$$

So (extracting subsequences) v_{ε} converges to v_o weakly in $\mathbb{L}^2(\Omega)$, and y_{ε} converges to y_o weakly in $\mathbb{L}^2(\Omega)$. It is easy to see that $y_o = T(v_o)$, $y_o \in K$, and $v_o \in U_{\text{ad}}$. Moreover,

$$J(\bar{y}, \bar{v}) \geq \liminf_{\varepsilon \rightarrow 0} J_{\varepsilon}(y_{\varepsilon}, v_{\varepsilon}) \geq \liminf_{\varepsilon \rightarrow 0} J(y_{\varepsilon}, v_{\varepsilon}) \geq J(y_o, v_o)$$

and

$$J(y_{\varepsilon}, v_{\varepsilon}) \leq J(\bar{y}, \bar{v}) \Rightarrow \liminf_{\varepsilon \rightarrow 0} J(y_{\varepsilon}, v_{\varepsilon}) = J(y_o, v_o) \leq J(\bar{y}, \bar{v}).$$

Therefore, $J(y_o, v_o) = J(\bar{y}, \bar{v})$ and $y_o = \bar{y}$, $v_o = \bar{v}$ (because the optimal solution of (Π) is unique). So we have just proved the weak convergence of y_{ε} to \bar{y} in $\mathbb{L}^2(\Omega)$, and v_{ε} to \bar{v} in $\mathbb{L}^2(\Omega)$.

Let us show the strong convergence; $\lim_{\varepsilon \rightarrow 0} J_{\varepsilon}(y_{\varepsilon}, v_{\varepsilon}) = J(\bar{y}, \bar{v})$ implies that

$$\lim_{\varepsilon \rightarrow 0} \|y_{\varepsilon} - \bar{y}\|_{\Omega}^2 + \|v_{\varepsilon} - \bar{v}\|_{\Omega}^2 = \|\bar{y} - \bar{y}\|_{\Omega}^2 + \|\bar{v} - \bar{v}\|_{\Omega}^2.$$

$\mathbb{L}^2(\Omega) \times \mathbb{L}^2(\Omega)$ is a Hilbert space, and so it is uniformly reflexive [11]. As we already have the weak convergence and the convergence of the norms, we then know that the convergence of v_{ε} to \bar{v} and y_{ε} to \bar{y} is strong in $\mathbb{L}^2(\Omega)$. Moreover, Ay_{ε} converges to $A\bar{y}$ strongly in $\mathbb{L}^2(\Omega)$, and A is an isomorphism; y_{ε} converges strongly to \bar{y} in $\mathbb{H}^2(\Omega) \cap \mathbb{H}_o^1(\Omega)$. \square

COROLLARY 3.4. *As ε tends to zero, p_ε converges to \bar{p} strongly in $\mathbb{H}^2(\Omega) \cap \mathbb{H}_o^1(\Omega)$, where \bar{p} is solution of*

$$(3.10) \quad \begin{aligned} A^* \bar{p} &= \bar{y} - z_d && \text{in } \Omega, \\ \bar{p} &= 0 && \text{on } \Gamma. \end{aligned}$$

COROLLARY 3.5. *There exists $\varepsilon_o > 0$, $k_o \geq 0$, such that*

$$\begin{aligned} \sup_{0 < \varepsilon \leq \varepsilon_o} \|y_\varepsilon\|_{\mathbb{H}^2(\Omega)} &\leq k_o; & \sup_{0 < \varepsilon \leq \varepsilon_o} \|v_\varepsilon\|_{\mathbb{L}^2(\Omega)} &\leq k_o, \\ \sup_{0 < \varepsilon \leq \varepsilon_o} \|\sqrt{\varepsilon} q_\varepsilon\|_{\mathbb{L}^2(\Omega)} &\leq k_o; & \sup_{0 < \varepsilon \leq \varepsilon_o} \|p_\varepsilon\|_{\mathbb{H}^2(\Omega)} &\leq k_o. \end{aligned}$$

3.1.2. Estimation of the multiplier q_ε . We must now estimate the multiplier q_ε to pass to the limit in the penalized optimality system (3.8), (3.9). So we must state an assumption stronger than (2.5) that allows us to obtain such an estimation.

The assumption we set shows that the interior of the convex \mathcal{D} must be nonempty (as does the Slater condition). We cannot avoid this kind of assumption, but we see that we may weaken it as far as possible if we consider the interior of \mathcal{D} for a stronger topology than the one of the state space Y .

Let \mathcal{E} be a dense subset of the control space $\mathbb{L}^2(\Omega)$ such that the injection is continuous. Let $\mathcal{B}(\mathcal{E})$ be the unit ball of \mathcal{E} : $\mathcal{B}(\mathcal{E}) = \{\kappa \in \mathcal{E} \mid \|\kappa\|_{\mathcal{E}} \leq 1\}$ ($\|\cdot\|_{\mathcal{E}}$ is the norm of \mathcal{E}).

We state the following assumption:

$$(3.11) \quad \begin{aligned} &\exists v_o \in U_{\text{ad}}, \quad \exists \rho > 0, \quad \exists R > 0 \text{ such that} \\ &\forall \kappa \in \mathcal{B}(\mathcal{E}), \quad \exists v_\kappa \in \mathcal{B}^2(v_o, R) \cap U_{\text{ad}}, \quad y_\kappa = T(f + v_\kappa - \rho\kappa) \in K. \end{aligned}$$

Remark 4. The condition $v_\kappa \in \mathcal{B}^2(v_o, R) \cap U_{\text{ad}}$ is not useful, of course, if U_{ad} is bounded (and may be replaced by $v_\kappa \in U_{\text{ad}}$). If U_{ad} is not bounded, we may suppose that v_o is given with (2.5) or $v_o = \bar{v}$. ($\mathcal{B}^2(v_o, R)$ is the \mathbb{L}^2 -ball, centered in v_o , of radius R .)

This assumption allows us to get an estimation of q_ε . We detail in the forthcoming section the optimality conditions to which different choices of \mathcal{E} lead.

PROPOSITION 3.6. *Let us assume (3.11). Then there exists $C \geq 0$, $\varepsilon_o > 0$ such that*

$$\sup_{0 < \varepsilon \leq \varepsilon_o} \|q_\varepsilon\|_{\mathcal{E}'} \leq k_o$$

(q_ε is bounded in the dual space of \mathcal{E} : \mathcal{E}').

Proof. Let κ belong to $\mathcal{B}(\mathcal{E})$. Let us add (3.8) and (3.9) applied to the couple (y_κ, v_κ) of $K \times U_{\text{ad}}$ defined by (3.11), as follows:

$$\begin{aligned} &\int_{\Omega} [q_\varepsilon + p_\varepsilon][A(y_\kappa - y_\varepsilon)] dx + \int_{\Omega} (Nv_\varepsilon - q_\varepsilon)(v_\kappa - v_\varepsilon) dx \geq 0, \\ &\int_{\Omega} p_\varepsilon(Ay_\kappa - f - v_\varepsilon) dx - \int_{\Omega} p_\varepsilon(Ay_\varepsilon - f - v_\varepsilon) dx + \int_{\Omega} q_\varepsilon(Ay_\kappa - v_\kappa - Ay_\varepsilon + v_\varepsilon) dx \\ &\quad + \int_{\Omega} Nv_\varepsilon(v_\kappa - v_\varepsilon) dx \geq 0, \\ &\int_{\Omega} p_\varepsilon(v_\kappa - v_\varepsilon - \rho\kappa) dx - \int_{\Omega} \varepsilon p_\varepsilon q_\varepsilon dx + \int_{\Omega} q_\varepsilon(\rho\kappa - \varepsilon q_\varepsilon) dx + \int_{\Omega} Nv_\varepsilon(v_\kappa - v_\varepsilon) dx \geq 0, \\ &\int_{\Omega} \rho q_\varepsilon \kappa dx \leq -\|\sqrt{\varepsilon} q_\varepsilon\|_{\Omega}^2 + \|Nv_\varepsilon\|_{\Omega} \|v_\kappa - v_\varepsilon\|_{\Omega} + \sqrt{\varepsilon} \|p_\varepsilon\|_{\Omega} \|\sqrt{\varepsilon} q_\varepsilon\|_{\Omega} \\ &\quad + \|p_\varepsilon\|_{\Omega} (\|v_\kappa - v_\varepsilon\|_{\Omega} + \rho \|\kappa\|_{\Omega}). \end{aligned}$$

Using Corollary 3.5, we have, for all $\varepsilon \in]0, \varepsilon_o]$,

$$\int_{\Omega} \rho q_{\varepsilon} \kappa \, dx \leq N k_o (\|v_{\kappa}\|_{\Omega} + k_o) + \sqrt{\varepsilon} k_o^2 + k_o (\|v_{\kappa}\|_{\Omega} + k_o + \rho \|\kappa\|_{\Omega}).$$

Moreover, the injection of \mathcal{E} in $\mathbb{L}^2(\Omega)$ is continuous so that

$$\forall \kappa \in \mathcal{B}(\mathcal{E}) \quad \|\kappa\|_{\Omega} \leq k_1 \|\kappa\|_{\mathcal{E}} = k_1.$$

Finally, as v_o does not depend on ε , and $\|v_{\kappa}\|_{\Omega} \leq (\|v_o\|_{\Omega} + R)$, we get

$$(3.12) \quad \exists C \geq 0, \quad \forall \varepsilon \in]0, \varepsilon_o], \quad \forall \kappa \in \mathcal{B}(\mathcal{E}) \quad \int_{\Omega} q_{\varepsilon} \kappa \, dx \leq C.$$

So q_{ε} is bounded in \mathcal{E}' . \square

We now specify what happens when we choose two different norms in \mathcal{E} .

3.1.3. A “strong” condition. Let \mathcal{E} be $\mathbb{L}^2(\Omega)$. Equation (3.11) then becomes

$$(3.13) \quad \begin{aligned} &\exists v_o \in U_{\text{ad}}, \quad \exists \rho > 0, \quad \exists R > 0 \quad \text{such that} \\ &\forall \kappa \in \mathcal{B}^2(\Omega), \quad \exists v_{\kappa} \in \mathcal{B}^2(v_o, R) \cap U_{\text{ad}}, \quad y_{\kappa} = T(f + v_{\kappa} - \rho \kappa) \in K, \end{aligned}$$

where $\mathcal{B}^2(\Omega)$ is the unit ball of $\mathbb{L}^2(\Omega)$.

We obtain the following result.

THEOREM 3.7. *Let us assume (3.13). Let (\bar{y}, \bar{v}) belong to \mathcal{D} , and \bar{p} be defined by (3.10). (\bar{y}, \bar{v}) is the optimal solution of (II) if and only if there exists $\bar{q} \in \mathbb{L}^2(\Omega)$ such that*

$$(3.13) \quad \forall y \in K \quad \int_{\Omega} (\bar{q} + \bar{p}) [A(y - \bar{y})] \, dx \geq 0,$$

$$(3.14) \quad \forall v \in U_{\text{ad}} \quad \int_{\Omega} (N\bar{v} - \bar{q})(v - \bar{v}) \, dx \geq 0.$$

Proof. Let us assume (3.13). Proposition 3.6 shows that q_{ε} is bounded in $\mathbb{L}^2(\Omega)$ (for $\mathbb{L}^2(\Omega)$ is its own dual).

Let (\bar{y}, \bar{v}) be the optimal solution of (II). q_{ε} is uniformly bounded (for the \mathbb{L}^2 -norm). We may then extract a subsequence that converges weakly to \bar{q} in $\mathbb{L}^2(\Omega)$. So, when ε tends to zero, the weak convergence in (3.8) and (3.9) gives us (3.13) and (3.14).

Conversely, let us choose $(y, v) \in \mathcal{D}$ and sum (3.13) and (3.14). As $Ay = f + v$ in Ω and $A\bar{y} = f + \bar{v}$ in Ω , we have

$$\int_{\Omega} (\bar{p} + N\bar{v})(v - \bar{v}) \, dx = \int_{\Omega} \bar{p} [A(y - \bar{y})] \, dx + \int_{\Omega} N\bar{v}(v - \bar{v}) \, dx \geq 0,$$

i.e.,

$$\forall (y, v) \in \mathcal{D}, \quad J'(\bar{y}, \bar{v})(y - \bar{y}, v - \bar{v}) \geq 0.$$

Thus (\bar{y}, \bar{v}) is the (unique) optimal solution of (II). \square

Remark 5. We have chosen $\mathcal{E} = \mathbb{L}^2(\Omega)$, but the proof is the same for any dense subset of $\mathbb{L}^2(\Omega)$. We may also choose very regular perturbations of the control setting, for instance, $\mathcal{E} = \mathcal{C}_o^{\infty}(\Omega)$ (the \mathcal{C}^{∞} -functions on Ω equal zero on Γ).

Remark 6. When $\mathcal{E} = \mathbb{L}^2(\Omega)$, (3.11) is equivalent to

$$\exists v_o \in U_{\text{ad}} \text{ such that } y_o = T(v_o) \in \text{Int}_{\mathbb{H}^2(\Omega)}(K)$$

($\text{Int}_X(K)$ is the interior of K in the sense of X -norm).

If $n \leq 3$, $\mathbb{H}^2(\Omega) \subset \mathcal{C}(\bar{\Omega})$, and (3.11) is equivalent to the Slater condition,

$$(3.15) \quad \exists u \in U_{\text{ad}}, \quad T(u) \in \text{Int}_{\mathbb{L}^{\infty}}(K).$$

So (according to Bonnans and Casas) if we choose U_{ad} such that $0 \in U_{\text{ad}}$ and $K = \{y \in \mathbb{H}^2(\Omega) \cap \mathbb{H}_o^1(\Omega) \mid |y(x)| \leq 1 \text{ in } \Omega\}$, (\mathcal{H}) is satisfied and Theorem 3.7 gives the optimality conditions.

Now let us consider another example. Let K and U_{ad} be the following sets ($n \leq 3$):

$$(3.15) \quad K = \{y \in \mathbb{H}^2(\Omega) \cap \mathbb{H}_o^1(\Omega) \mid \varphi(x) \leq y(x) \leq \psi(x) \text{ a.e. on } \Omega\},$$

$$(3.16) \quad U_{\text{ad}} = \{v \in \mathbb{L}^2(\Omega) \mid \alpha(x) \leq v(x) \leq \beta(x)\} \text{ in } \Omega.$$

If $\varphi|_{\Gamma} = \psi|_{\Gamma}$, the Slater condition (i.e., (\mathcal{H})) cannot be satisfied because the \mathbb{L}^∞ -interior of K is empty. So we would like now to weaken assumption (3.11) to extend the results of Theorem 3.7. Of course, under weaker assumptions we obtain a “weaker” optimality system. That will be done by changing the space \mathcal{E} and, more precisely, the topology of \mathcal{E} .

3.1.4. A “weak” condition. Let \mathcal{E} be $\mathcal{C}_o(\Omega)$ with the \mathbb{L}^∞ -norm. ($\mathcal{C}_o(\Omega)$ is the set of continuous functions equal to zero on Γ .) Equation (3.11) then becomes

$$(\mathcal{H}^*) \quad \begin{aligned} &\exists v_o \in U_{\text{ad}}, \quad \exists \rho > 0, \quad \exists R > 0 \quad \text{such that} \\ &\forall \kappa \in \mathcal{B}^\infty(\mathcal{C}_o(\Omega)), \quad \exists v_\kappa \in \mathcal{B}^2(v_o, R) \cap U_{\text{ad}}, \quad y_\kappa = T(f + v_\kappa - \rho \kappa) \in K, \end{aligned}$$

($\mathcal{B}^\infty(\mathcal{C}_o(\Omega)) = \{\kappa \in \mathcal{C}_o(\Omega) \mid \|\kappa\|_{\mathbb{L}^\infty} \leq 1\}$ is the unit ball of $\mathcal{C}_o(\Omega)$).

Remark 7. Let us consider K and U_{ad} defined with (3.15) and (3.16), and assume that we can find

1. $e \in \mathcal{C}_o^1(\Omega)$ such that $e > 0$ in Ω , $\partial e / \partial n < 0$, and $\varphi + e \leq \psi - e$ a.e. in Ω ;
2. $v_o \in U_{\text{ad}}$ such that $y_o = T(v_o) \in [\varphi + e, \psi - e]$;

then (\mathcal{H}^*) is satisfied.

We first note that it is sufficient to find $\rho > 0$ such that for all $\kappa \in \mathcal{B}^\infty(\mathcal{C}_o(\Omega))$, $x \in \Omega$,

$$(i) \quad |z(x)| - \frac{e(x)}{\rho} \leq 0,$$

where $z = T(\kappa)$.

Let κ be in $\mathcal{B}^\infty(\mathcal{C}_o(\Omega))$. $z = T(\kappa) \in \mathcal{C}_o^2(\Omega)$ and $\|\nabla z\|_\infty \leq M$, $\|z\|_\infty \leq M$, where M is independent of κ . Moreover, as $e \in \mathcal{C}_o^1(\Omega)$, we can find $m > 0$ such that for all $x \in \Gamma$, $(\partial e / \partial n)(x) \leq -m < 0$.

We first show that (i) is true (with $\rho \geq m/(2M)$) on an open subset \mathcal{V} of Ω , such that $\Gamma \subset \bar{\mathcal{V}}$.

Let x belong to Γ , and ν_x the exterior normal to Γ at x . We can see that (i) is satisfied on $I_x =]x, x - \alpha_x \nu_x[$, where α_x is small enough and nonnegative. (We use the Taylor formula applied to the restriction of e to I_x .) Then we set $\mathcal{V} = \bigcup \{I(x) \mid x \in \Gamma\}$. (We use the regularity properties of Ω and Γ , and the fact that Ω is locally on the same side of Γ .)

Then we may find a compact subset C of Ω such that $C \cap \Gamma = \emptyset$. As the function e is continuous on C and strictly positive, we can find $\tilde{m} > 0$ such that $e(x) \geq \tilde{m}$ for each x of C . So (i) is satisfied on C if we choose $\rho \geq \tilde{m}/M$.

Remark 8. We may note that (\mathcal{S}) (defined with Remark 6) implies (\mathcal{H}^*) , but is not equivalent because T is no longer an isomorphism from $\mathbb{L}^\infty(\Omega)$ to $\mathbb{H}^2(\Omega) \cap \mathbb{H}_o^1(\Omega)$.

Proposition 3.6 ensures that q_e is bounded in the dual space of \mathcal{E} , i.e., in the space of Radon measures on Ω : $\mathcal{M}(\Omega)$. As before, we would like to pass to the limit in the penalized optimality system. The main difficulty is that we must choose test functions

smooth enough to take the limit, in the dual sense $\langle \mathcal{M}(\Omega), \mathcal{C}_o(\Omega) \rangle$. So we define

$$(3.17) \quad \mathcal{F}(\bar{y}) = \{y \in \mathbb{H}^2(\Omega) \mid \exists \varphi \in \mathcal{C}_o(\Omega), Ay = A\bar{y} + \varphi \text{ a.e. on } \Omega\},$$

$$(3.18) \quad \mathcal{G}(\bar{v}) = \bar{v} + \mathcal{C}_o(\Omega).$$

(We note that $\bar{y} \in \mathcal{F}(\bar{y})$ and $\bar{v} \in \mathcal{G}(\bar{v})$).

We obtain the following result.

THEOREM 3.8. *Let us assume (\mathcal{H}^*) and let (\bar{y}, \bar{v}) be the solution of (II). Then there exists $\bar{q} \in \mathcal{M}(\Omega)$ such that*

$$(3.19) \quad \forall y \in K \cap \mathcal{F}(\bar{y}) \quad \int_{\Omega} (\bar{q} + \bar{p})[A(y - \bar{y})] dx \geq 0,$$

$$(3.20) \quad \forall v \in U_{\text{ad}} \cap \mathcal{G}(\bar{v}) \quad \int_{\Omega} (N\bar{v} - \bar{q})(v - \bar{v}) dx \geq 0.$$

Proof. We know (Proposition 3.6) that q_ε is uniformly bounded in $\mathcal{M}(\Omega)$, and we can then extract a subsequence that converges to \bar{q} in $\mathcal{M}(\Omega)$ (with weak star topology).

We cannot pass to the limit immediately in the penalized optimality system because Ay_ε does not converge to $A\bar{y}$ in $\mathcal{C}_o(\Omega)$. (We do not even know if Ay_ε belongs to $\mathcal{C}_o(\Omega)$ or if $A\bar{y}$ is in $\mathcal{C}_o(\Omega)$.) Nevertheless, we may write (3.8) and (3.9) as the following:

$$\forall y \in K \quad \int_{\Omega} p_\varepsilon A(y - y_\varepsilon) dx + \int_{\Omega} q_\varepsilon [A(y - \bar{y}) + (A\bar{y} - f) - (Ay_\varepsilon - f)] dx \geq 0,$$

$$\forall v \in U_{\text{ad}} \quad \int_{\Omega} Nv_\varepsilon(v - v_\varepsilon) dx - \int_{\Omega} q_\varepsilon [(v - \bar{v}) + (\bar{v} - v_\varepsilon)] dx \geq 0,$$

i.e.,

$$\forall y \in K \quad \int_{\Omega} p_\varepsilon A(y - y_\varepsilon) dx + \int_{\Omega} q_\varepsilon A(y - \bar{y}) dx + \int_{\Omega} q_\varepsilon (\bar{v} - v_\varepsilon) dx - \varepsilon \|q_\varepsilon\|_{\Omega}^2 \geq 0,$$

$$\forall v \in U_{\text{ad}} \quad \int_{\Omega} Nv_\varepsilon(v - v_\varepsilon) dx - \int_{\Omega} q_\varepsilon (v - \bar{v}) dx - \int_{\Omega} q_\varepsilon (\bar{v} - v_\varepsilon) dx \geq 0.$$

Adding these two inequalities, we obtain for all $y \in K$, $v \in U_{\text{ad}}$,

$$\int_{\Omega} p_\varepsilon A(y - y_\varepsilon) dx + \int_{\Omega} q_\varepsilon A(y - \bar{y}) dx + \int_{\Omega} Nv_\varepsilon(v - v_\varepsilon) dx - \int_{\Omega} q_\varepsilon (v - \bar{v}) dx \geq 0.$$

We know (without needing (\mathcal{H}^*)) that for all $y \in \mathbb{H}^2(\Omega)$, $v \in \mathbb{L}^2(\Omega)$,

$$\int_{\Omega} p_\varepsilon A(y - y_\varepsilon) dx \rightarrow \int_{\Omega} \bar{p} A(y - \bar{y}) dx \quad \text{and} \quad \int_{\Omega} Nv_\varepsilon(v - v_\varepsilon) dx \rightarrow \int_{\Omega} N\bar{v}(v - \bar{v}) dx.$$

Now we can pass to the limit in $\int_{\Omega} q_\varepsilon A(y - \bar{y}) dx$ and $\int_{\Omega} q_\varepsilon (v - \bar{v}) dx$ if and only if $A(y - \bar{y})$ and $(v - \bar{v})$ are “smooth enough,” i.e., $y \in \mathcal{F}(\bar{y})$ and $v \in \mathcal{G}(\bar{v})$. Thus we get for all $y \in K \cap \mathcal{F}(\bar{y})$, $v \in U_{\text{ad}} \cap \mathcal{G}(\bar{v})$,

$$\int_{\Omega} [\bar{p} + \bar{q}][A(y - \bar{y})] dx + \int_{\Omega} [N\bar{v} - \bar{q}][v - \bar{v}] dx \geq 0.$$

Then, decoupling again (with the successive choices of $y = \bar{y}$ and $v = \bar{v}$), we obtain (3.19) and (3.20). \square

Remark 9. As the test functions must be smooth we lose, of course, the fact that the optimality system is sufficient to ensure that (\bar{y}, \bar{v}) is the optimal solution. Moreover, the difficulty still remaining is to be sure that the test functions spaces $K \cap \mathcal{F}(\bar{y})$ and $U_{\text{ad}} \cap \mathcal{G}(\bar{v})$ are large enough to allow good numerical experimentation. It is the case, for example, if \bar{v} and f belongs to $\mathbb{H}^r(\Omega)$ with $r > n/2$: $\bar{v} \in \mathcal{C}^o(\bar{\Omega})$ and $\bar{y} \in \mathbb{H}^{r+2}(\Omega) \subseteq \mathcal{C}^2(\bar{\Omega})$.

Remark 10. Let us consider the particular case where $A = -\Delta + \text{Id}$, $f = 0$, $U_{\text{ad}} = \mathbb{L}^2(\Omega)$, and $K = \{y \in \mathbb{H}^2(\Omega) \cap \mathbb{H}_o^1(\Omega) \mid |y| \leq \alpha\}$. This problem is not qualified and has been studied by Abergel and Temam [2], who have proved the existence of an adjoint state (or multiplier) in $BL_o(\Omega)$, where $BL_o(\Omega) = \{q \in \mathbb{L}^2(\Omega) \mid -\Delta q + q \in \mathcal{M}(\Omega), q = 0 \text{ on } \Gamma\}$ ($\mathcal{M}(\Omega)$ bounded measures on Ω), and established sufficient and necessary optimality conditions for \bar{y} to be an optimal solution. Let us apply the penalization to the following problem. Proposition 3.2 gives

$$(a) \quad \forall y \in K \quad \int_{\Omega} [p_{\varepsilon} + q_{\varepsilon}] [-\Delta(y - y_{\varepsilon}) + (y - y_{\varepsilon})] dx \geq 0,$$

$$(b) \quad \forall v \in \mathbb{L}^2(\Omega) \quad \int_{\Omega} (Nv_{\varepsilon} - q_{\varepsilon})(v - v_{\varepsilon}) dx \geq 0.$$

Part (b) means that $q_{\varepsilon} = Nv_{\varepsilon}$ and so q_{ε} converges strongly in $\mathbb{L}^2(\Omega)$ to $\bar{q} = N\bar{v}$. We may pass to the limit in (a) without any qualification hypothesis, and we get

$$(c) \quad \forall y \in K \quad \int_{\Omega} [\bar{p} + \bar{q}] [-\Delta(y - \bar{y}) + (y - \bar{y})] dx \geq 0.$$

So if $\bar{q} \in BL_o(\Omega)$, we may define the measure $\bar{r} = -\Delta\bar{q} + \bar{q}$ according to [2], and

$$\forall y \in \mathbb{H}^2(\Omega) \cap \mathbb{H}_o^1(\Omega) \quad \int_{\Omega} (-\Delta\bar{q} + \bar{q})y = \int_{\Omega} \bar{q}(-\Delta y + y) dx.$$

Finally, \bar{y} is an optimal solution of (II) if and only if there exists, in $BL_o(\Omega)$, a multiplier (or adjoint state) \bar{q} such that

$$(d) \quad \forall y \in K \quad \int_{\Omega} [\bar{y} - z_d - \Delta\bar{q} + \bar{q}][y - \bar{y}] \geq 0, \\ \bar{q} = N(-\Delta\bar{y} + \bar{y}).$$

Let us note that

$$(d) \Leftrightarrow \int_{\Omega} (-z_d - \Delta\bar{q} + \bar{q})\bar{y} + \int_{\Omega} \bar{y}^2 dx = \inf_{y \in K} \left\{ \int_{\Omega} (\bar{y} - z_d - \Delta\bar{q} + \bar{q})(y) \right\}.$$

So we obtain an optimality system similar to the one of Temam and Abergel. Nevertheless, we are not able to prove that \bar{q} belongs to $BL_o(\Omega)$ with a proof different from the one of the above authors.

Remark 11. The weak optimality system ensures the existence of Lagrange multipliers (as measures) without the Slater condition. We may then use, for instance, Lagrangian methods to compute the solution. These methods are min-max methods dealing with the Lagrangian of the problem, which is well defined because the multiplier exists (cf. [15]).

3.2. Neumann boundary conditions.

3.2.1. Penalization. The method is the same as in § 3.1, so we just give results without detailed proofs (for more details, see [5]).

We now consider the following Neumann problem:

$$(3.21) \quad \mathbf{P}(v) \begin{cases} Ay = f + v & \text{sur } \Omega, \\ \frac{\partial y}{\partial v_A} = 0 & \text{sur } \Gamma, \end{cases}$$

where $v \in \mathbb{L}^2(\Omega)$, $f \in \mathbb{L}^2(\Omega)$.

$\mathbf{P}(v)$ has a unique solution $T(v)$ in $\mathbb{H}^2(\Omega)$ (cf. [19]) and $T: v \mapsto y = T(v)$ is affine. Let K be a nonempty, convex, closed subset of $\mathbb{H}^2(\Omega)$,

$$\left(y \in \mathbb{H}^2(\Omega) \Rightarrow \frac{\partial y}{\partial v_A} \in \mathbb{H}^{1/2}(\Gamma) \subseteq \mathbb{L}^2(\Gamma) \right).$$

Let U_{ad} be a nonempty, closed, convex subset of $\mathbb{L}^2(\Omega)$. For $\varepsilon > 0$, let us define J_ε on $\mathbb{V} \times \mathbb{L}^2(\Omega)$ by

$$(3.22) \quad J_\varepsilon(y, v) = J(y, v) + \frac{1}{2\varepsilon} \int_{\Omega} (Ay - f - v)^2 dx + \frac{1}{2\varepsilon} \int_{\Gamma} \left(\frac{\partial y}{\partial v} \right)_A^2 d\Gamma.$$

PROPOSITION 3.9. (Π_ε) has a unique solution $(y_\varepsilon, v_\varepsilon)$ in $K \times U_{\text{ad}}$ and

$$\forall (y, v) \in K \times U_{\text{ad}}, \quad J'_\varepsilon(y_\varepsilon, v_\varepsilon)(y - y_\varepsilon, v - v_\varepsilon) \geq 0.$$

So we call

$$(3.23) \quad \left[\begin{array}{l} q_\varepsilon = \frac{Ay_\varepsilon - f - v_\varepsilon}{\varepsilon} \in \mathbb{L}^2(\Omega), \\ r_\varepsilon = \frac{1}{\varepsilon} \frac{\partial y_\varepsilon}{\partial v_A} \in \mathbb{L}^2(\Gamma), \\ \text{and let } p_\varepsilon \in \mathbb{H}^2(\Omega) \text{ be the adjoint state solution of} \\ A^* p_\varepsilon = y_\varepsilon - z_d \quad \text{in } \Omega, \\ \frac{\partial p_\varepsilon}{\partial v_A} = 0 \quad \text{on } \Gamma. \end{array} \right.$$

The penalized optimality system is given by the following result.

PROPOSITION 3.10. Let $(y_\varepsilon, v_\varepsilon)$ be the optimal solution of (Π_ε) . Then

$$(3.24) \quad \forall y \in K \quad \int_{\Gamma} (r_\varepsilon + p_\varepsilon) \frac{\partial(y - y_\varepsilon)}{\partial v_A} d\Gamma + \int_{\Omega} [q_\varepsilon + p_\varepsilon][A(y - y_\varepsilon)] dx \geq 0,$$

$$(3.25) \quad \forall v \in U_{\text{ad}} \quad \int_{\Omega} (Nv_\varepsilon - q_\varepsilon)(v - v_\varepsilon) dx \geq 0.$$

The following theorem describes the asymptotic behaviour of y_ε , v_ε , and p_ε .

THEOREM 3.11. When ε tends to 0,

- (i) v_ε converges to \bar{v} strongly in $\mathbb{L}^2(\Omega)$;
- (ii) y_ε converges to \bar{y} strongly in $\mathbb{H}^2(\Omega)$, where (\bar{y}, \bar{v}) is the solution of (Π) ;
- (iii) p_ε converges to \bar{p} strongly in $\mathbb{H}^2(\Omega)$, where \bar{p} is the solution of

$$(3.26) \quad \begin{cases} A^* \bar{p} = \bar{y} - z_d & \text{in } \Omega, \\ \frac{\partial \bar{p}}{\partial v_A} = 0 & \text{on } \Gamma; \end{cases}$$

- (iv) $\partial y_\varepsilon / \partial v_A$ converges to $\partial \bar{y} / \partial v_A = 0$ (strongly in $\mathbb{L}^2(\Gamma)$).

Now we must estimate the two multipliers, q_ε and r_ε . So let \mathcal{E}_Ω and \mathcal{E}_Γ be two dense subsets of, respectively, $\mathbb{L}^2(\Omega)$ and $\mathbb{L}^2(\Gamma)$ such that the injections are continuous. We state the following assumption:

$$(3.27) \quad \begin{aligned} & \exists \rho > 0, \quad \exists v_o \in U_{\text{ad}}, \quad \exists R > 0 \quad \text{such that} \\ & \forall \eta = (\kappa, \xi) \in \mathcal{B}(\mathcal{E}_\Omega) \times \mathcal{B}(\mathcal{E}_\Gamma), \quad \exists v_\eta \in \mathcal{B}^2(v_o, R) \cap U_{\text{ad}} \quad \text{such that } y_\eta \in K, \end{aligned}$$

where y_η is the solution of

$$(P_\eta) \quad \begin{aligned} & Ay_\eta = f + v_\eta - \rho \kappa \quad \text{in } \Omega, \\ & \frac{\partial y_\eta}{\partial v_A} = -\rho \xi \quad \text{on } \Gamma. \end{aligned}$$

We may then ensure that r_ε and q_ε are uniformly bounded in the dual spaces of \mathcal{E}_Ω and \mathcal{E}_Γ .

PROPOSITION 3.12. *Assuming (3.27), there exists $C > 0$, $\varepsilon_o > 0$ such that*

$$\forall \varepsilon \in]0, \varepsilon_o], \quad \forall \kappa \in \mathcal{B}(\mathcal{E}_\Omega), \quad \forall \xi \in \mathcal{B}(\mathcal{E}_\Gamma) \quad \int_{\Omega} q_\varepsilon \kappa \, dx + \int_{\Gamma} r_\varepsilon \xi \, d\Gamma \leq C.$$

Then we still have two different optimality systems more or less “strong.”

3.2.2. The “strong” optimality system. Let us choose (as before) $\mathcal{E}_\Omega = \mathbb{L}^2(\Omega)$ and $\mathcal{E}_\Gamma = \mathbb{L}^2(\Gamma)$.

THEOREM 3.13. *Let (\bar{y}, \bar{v}) belong to \mathcal{D} and let us assume*

$$(H) \quad \begin{aligned} & \exists \rho > 0, \quad \exists v_o \in U_{\text{ad}}, \quad \exists R > 0 \quad \text{such that} \\ & \forall \eta = (\kappa, \xi) \in \mathcal{B}^2(\Omega) \times \mathcal{B}^2(\Gamma), \quad \exists v_\eta \in \mathcal{B}^2(v_o, R) \cap U_{\text{ad}}, \\ & \text{such that the solution } y_\eta \text{ of } (P_\eta) \text{ belongs to } K. \end{aligned}$$

(\bar{y}, \bar{v}) is the optimal solution of (Π) if and only if there exists $\bar{q} \in \mathbb{L}^2(\Omega)$, $\bar{r} \in \mathbb{L}^2(\Gamma)$ such that

$$(3.28) \quad \forall y \in K \quad \int_{\Gamma} (\bar{r} + \bar{p}) \frac{\partial(y - \bar{y})}{\partial v_A} \, d\Gamma + \int_{\Omega} [\bar{q} + \bar{p}][A(y - \bar{y})] \, dx \geq 0,$$

$$(3.29) \quad \forall v \in U_{\text{ad}} \quad \int_{\Omega} (N\bar{v} - \bar{q})(v - \bar{v}) \, dx \geq 0,$$

where \bar{p} is the adjoint state given by (3.26).

Proof. The proof is exactly the same as that of Theorem 3.7. \square

3.2.3. The “weak” optimality system. It holds that $\mathcal{E}_\Omega = \mathcal{C}_o(\Omega)$ and $\mathcal{E}_\Gamma = \mathcal{C}_o(\Gamma)$ (with the \mathbb{L}^∞ -norm).

As before, the test functions must be smooth enough to get the convergence in the dual sense. So we set

$$\mathcal{F}(\bar{y}) = \left\{ y \in \mathbb{V} \mid \exists (\varphi, \psi) \in \mathcal{C}_o(\Omega) \times \mathcal{C}_o(\Gamma), \quad Ay = A\bar{y} + \varphi \text{ on } \Omega, \quad \frac{\partial y}{\partial v_A} = \frac{\partial \bar{y}}{\partial v_A} + \psi \text{ on } \Gamma \right\},$$

$$\mathcal{G}(\bar{v}) = \bar{v} + \mathcal{C}_o(\Omega),$$

and we obtain a “weak” optimality system, which is no more sufficient.

THEOREM 3.14. *Let (\bar{y}, \bar{v}) be the optimal solution of (π) and assume that*

$$(H^*) \quad \begin{aligned} & \exists \rho > 0, \quad \exists v_o \in U_{\text{ad}}, \quad \exists R > 0 \quad \text{such that} \\ & \forall \eta = (\kappa, \xi) \in \mathcal{B}^\infty(\mathcal{C}_o(\Omega)) \times \mathcal{B}^\infty(\mathcal{C}_o(\Gamma)), \quad \exists v_\eta \in \mathcal{B}^2(v_o, R) \cap U_{\text{ad}}, \\ & \text{such that the solution } y_\eta \text{ of } (P_\eta) \text{ belongs to } K. \end{aligned}$$

Then there exists $\bar{q} \in \mathcal{M}(\Omega)$, $\bar{r} \in \mathcal{M}(\Gamma)$ such that

$$(3.30) \quad \forall y \in K \cap \mathcal{F}(\bar{y}) \quad \int_{\Gamma} (\bar{r} + \bar{p}) \frac{\partial(y - \bar{y})}{\partial v_A} d\Gamma + \int_{\Omega} [\bar{q} + \bar{p}][A(y - \bar{y})] dx \geq 0,$$

$$(3.31) \quad \forall v \in U_{ad} \cap \mathcal{G}(N\bar{v}) \quad \int_{\Omega} (N\bar{v} - \bar{q})(v - \bar{v}) dx \geq 0,$$

where \bar{p} is the adjoint state given by (3.26).

4. Case II: Boundary control problems.

4.1. Dirichlet boundary conditions.

4.1.1. Penalization. Let us consider the following *Dirichlet* problem:

$$(4.1) \quad \mathbf{P}(v) \begin{cases} Ay = f & \text{in } \Omega, \\ y = v & \text{on } \Gamma, \end{cases}$$

where $f \in \mathbb{L}^2(\Omega)$ and $v \in \mathbb{L}^2(\Gamma) (\subseteq \mathbb{H}^{-1/2}(\Gamma))$.

$\mathbf{P}(v)$ has a unique solution $y = T(v) \in \mathbb{W} = \{y \in \mathbb{L}^2(\Omega) \mid Ay \in \mathbb{L}^2(\Omega)\}$ (with the norm $\|y\|_{\mathbb{W}} = \|y\|_{\Omega} + \|Ay\|_{\Omega}$), and T is affine (linear if $f=0$). K (respectively, U_{ad}) is a nonempty, closed, convex subset of \mathbb{W} (respectively, $\mathbb{L}^2(\Gamma)$).

Remark 12. If we consider pointwise constraints, we must be able to define $y(x)$ everywhere in $\bar{\Omega}$. If $n \leq 3$, it is sufficient (cf. [19]) to take $U_{ad} \subseteq \mathbb{H}^{3/2}(\Gamma)$ so that $y \in \mathbb{H}^2(\Omega) (\subseteq \mathcal{C}(\bar{\Omega}))$.

We consider, for example, the “cost” functional defined on $\mathbb{L}^2(\Omega) \times \mathbb{L}^2(\Gamma)$

$$J(y, v) = \frac{1}{2} \int_{\Omega} (y - z_d)^2 dx + \frac{N}{2} \int_{\Gamma} v^2 dx.$$

We always consider the control problem (Π) (cf. § 2) and assume that (2.5) is satisfied. $\varepsilon > 0$ being given, let us define J_{ε} on $\mathbb{W} \times \mathbb{L}^2(\Gamma)$ with

$$(4.2) \quad J_{\varepsilon}(y, v) = \begin{cases} J(y, v) + \frac{1}{2\varepsilon} \int_{\Gamma} (y - v)^2 d\Gamma + \frac{1}{2\varepsilon} \int_{\Omega} (Ay + f)^2 dx, & \text{if } y|_{\Gamma} \in \mathbb{L}^2(\Gamma), \\ +\infty & \text{else.} \end{cases}$$

Remark 13. The second penalization term does not come from a relation between the state y and the control v . Anyway, it is more interesting to introduce it to simplify the convex where the test functions y are to be chosen. (If we omit it, we must take y in $\mathcal{H} = \{y \in K \mid Ay = f \text{ in } \Omega\}$.) Moreover, if y belongs to \mathbb{W} we do not know whether the trace of y on Γ belongs to $\mathbb{L}^2(\Gamma)$. We know that this trace can be defined in $\mathbb{H}^{-1/2}(\Gamma)$. So we define K^* as $K^* = \{y \in K \mid y|_{\Gamma} \in \mathbb{L}^2(\Gamma)\}$.

PROPOSITION 4.1. (Π_{ε}) has a unique solution $(y_{\varepsilon}, v_{\varepsilon})$ and

$$\forall (y, v) \in K^* \times U_{ad}, \quad J'_{\varepsilon}(y_{\varepsilon}, v_{\varepsilon})(y - y_{\varepsilon}, v - v_{\varepsilon}) \geq 0.$$

So we get

$$\begin{aligned} \forall y \in K^* \quad \int_{\Omega} (y_{\varepsilon} - z_d)(y - y_{\varepsilon}) dx + \frac{1}{\varepsilon} \int_{\Omega} [Ay_{\varepsilon} - f][A(y - y_{\varepsilon})] dx \\ + \frac{1}{\varepsilon} \int_{\Gamma} (y_{\varepsilon} - v_{\varepsilon})(y - y_{\varepsilon}) d\Gamma \geq 0, \end{aligned}$$

$$\forall v \in U_{ad} \quad \int_{\Gamma} \left[Nv_{\varepsilon} - \frac{1}{\varepsilon} (y_{\varepsilon} - v_{\varepsilon}) \right] [v - v_{\varepsilon}] d\Gamma \geq 0.$$

Let us define

$$(4.3) \quad \begin{aligned} q_\varepsilon &= \frac{y_\varepsilon - v_\varepsilon}{\varepsilon} \in \mathbb{L}^2(\Gamma), \\ s_\varepsilon &= \frac{Ay_\varepsilon - f}{\varepsilon} \in \mathbb{L}^2(\Omega). \end{aligned}$$

Let $p_\varepsilon \in (\mathbb{H}_o^1(\Omega) \cap \mathbb{H}^2(\Omega))$ be the solution of the dual problem

$$(4.4) \quad \begin{aligned} A^*p_\varepsilon &= y_\varepsilon - z_d \quad \text{in } \Omega, \\ p_\varepsilon &= 0 \quad \text{on } \Gamma. \end{aligned}$$

Using a henceforth classical calculus, we obtain the following penalized optimality system.

PROPOSITION 4.2. *Let $(y_\varepsilon, v_\varepsilon)$ be the solution of (Π_ε) ; then*

$$(4.5) \quad \forall y \in K \quad \int_\Gamma \left(-\frac{\partial p_\varepsilon}{\partial v_{A^*}} + q_\varepsilon \right) (y - y_\varepsilon) d\Gamma + \int_\Omega [s_\varepsilon + p_\varepsilon][A(y - y_\varepsilon)] dx \geq 0,$$

$$(4.6) \quad \forall v \in U_{\text{ad}} \quad \int_\Gamma (Nv_\varepsilon - q_\varepsilon)(v - v_\varepsilon) d\Gamma \geq 0.$$

The first result about the asymptotic behaviour of ε -quantities is given by the theorem below.

THEOREM 4.3. *When ε tends to 0,*

- (i) v_ε converges to \bar{v} strongly in $\mathbb{L}^2(\Gamma)$;
- (ii) y_ε converges to \bar{y} strongly in \mathbb{W} ((\bar{y}, \bar{v}) is the solution of (Π));
- (iii) p_ε converges to \bar{p} strongly in $\mathbb{H}^2(\Omega) \cap \mathbb{H}_o^1(\Omega)$, where \bar{p} is the adjoint state solution of

$$(4.7) \quad \begin{aligned} A^*\bar{p} &= \bar{y} - z_d \quad \text{in } \Omega, \\ \bar{p} &= 0 \quad \text{on } \Gamma; \end{aligned}$$

- (iv) $\lim_{\varepsilon \rightarrow 0} \partial p_\varepsilon / \partial v_{A^*} = \partial \bar{p} / \partial v_{A^*}$ strongly in $\mathbb{H}^{1/2}(\Gamma)$.

As we must now estimate q_ε and s_ε , we assume once again that

$$(4.8) \quad \begin{aligned} &\exists \rho > 0, \quad \exists v_o \in U_{\text{ad}}, \quad \exists R > 0 \quad \text{such that} \\ &\forall \eta = (\kappa, \xi) \in \mathcal{B}(\mathcal{E}_\Omega) \times \mathcal{B}(\mathcal{E}_\Gamma), \quad \exists v_\eta \in \mathcal{B}^2(v_o, R) \cap U_{\text{ad}} \quad \text{such that } y_\eta \in K, \end{aligned}$$

where y_η is the solution of

$$(\mathbf{P}_\eta) \begin{cases} Ay_\eta = f - \rho\kappa & \text{in } \Omega, \\ y_\eta = v_\eta - \rho\xi & \text{on } \Gamma, \end{cases}$$

where \mathcal{E}_Ω and \mathcal{E}_Γ are dense subsets of, respectively, $\mathbb{L}^2(\Omega)$ and $\mathbb{L}^2(\Gamma)$, such that the injections are continuous and $\mathcal{B}(\mathcal{E})$ is the unit ball of \mathcal{E} (for the \mathcal{E} -norm).

Then q_ε (respectively, s_ε) is bounded in the dual space of \mathcal{E}_Γ (respectively, \mathcal{E}_Ω).

PROPOSITION 4.4. *Assuming (4.8), there exists $C > 0$, $\varepsilon_o > 0$ such that*

$$\forall \varepsilon \in]0, \varepsilon_o], \quad \forall (\kappa, \xi) \in \mathcal{B}(\mathcal{E}_\Omega) \times \mathcal{B}(\mathcal{E}_\Gamma) \quad \int_\Gamma q_\varepsilon \xi d\Gamma + \int_\Omega s_\varepsilon \kappa dx \leq C.$$

Proof. The proof is the same as that of Proposition 3.6, but we are going to detail it somewhat.

Let (κ, ξ) belong to $\mathcal{B}(\mathcal{E}_\Omega) \times \mathcal{B}(\mathcal{E}_\Gamma)$. Let us sum (4.5) and (4.6) applied to the couple (y_η, v_η) of $K^* \times U_{\text{ad}}$ given by (4.8). We obtain

$$\begin{aligned} & \int_{\Gamma} \left(-\frac{\partial p_\varepsilon}{\partial v_{A^*}} + q_\varepsilon \right) (y_\eta - y_\varepsilon) d\Gamma + \int_{\Omega} [s_\varepsilon + p_\varepsilon][A(y_\eta - y_\varepsilon)] dx \\ & \quad + \int_{\Gamma} (Nv_\varepsilon - q_\varepsilon)(v_\eta - v_\varepsilon) d\Gamma \geq 0, \\ & \int_{\Gamma} \left(-\frac{\partial p_\varepsilon}{\partial v_{A^*}} + q_\varepsilon \right) (v_\eta - v_\varepsilon - p\xi - \varepsilon q_\varepsilon) d\Gamma + \int_{\Omega} [s_\varepsilon + p_\varepsilon][-\rho\kappa - \varepsilon s_\varepsilon] dx \\ & \quad + \int_{\Gamma} (Nv_\varepsilon - q_\varepsilon)(v_\eta - v_\varepsilon) d\Gamma \geq 0, \\ & \int_{\Omega} \rho(s_\varepsilon\kappa + q_\varepsilon\xi) dx \leq \int_{\Gamma} \left(Nv_\varepsilon - \frac{\partial p_\varepsilon}{\partial v_{A^*}} \right) (v_\eta - v_\varepsilon) d\Gamma + \int_{\Omega} p_\varepsilon(-\rho\kappa - \varepsilon s_\varepsilon) dx \\ & \quad + \int_{\Gamma} \left(-\frac{\partial p_\varepsilon}{\partial v_{A^*}} \right) (-\rho\xi - \varepsilon q_\varepsilon) d\Gamma - \varepsilon \int_{\Gamma} (q_\varepsilon^2 + s_\varepsilon^2) d\Gamma. \end{aligned}$$

As before, all the quantities of the right side are bounded, and we have the wanted inequality. \square

4.1.2. The optimality system. As before, we make two different choices for the norm of \mathcal{E} . First, we choose $\mathcal{E}_\Omega = \mathbb{L}^2(\Omega)$ and $\mathcal{E}_\Gamma = \mathbb{L}^2(\Gamma)$, so that (4.8) becomes

$$\exists \rho > 0, \quad \exists v_o \in U_{\text{ad}}, \quad \exists R > 0 \quad \text{such that}$$

$$(\mathcal{H}) \quad \forall \eta = (\kappa, \xi) \in \mathcal{B}^2(\Omega) \times \mathcal{B}^2(\Gamma), \quad \exists v_\eta \in \mathcal{B}^2(v_o, R) \cap U_{\text{ad}} \quad \text{such that}$$

the solution y_η of (P_η) belongs to K .

We obtain a “strong” optimality system, shown by the following result.

THEOREM 4.5. *Let us assume (\mathcal{H}) ; let (\bar{y}, \bar{v}) be in \mathcal{D} and \bar{p} be defined with (4.7). (\bar{y}, \bar{v}) is the optimal solution of (Π) if and only if there exists $\bar{q} \in \mathbb{L}^2(\Gamma)$, $\bar{s} \in \mathbb{L}^2(\Omega)$ such that*

$$(4.9) \quad \forall y \in K^* \quad \int_{\Gamma} \left(-\frac{\partial \bar{p}}{\partial v_{A^*}} + \bar{q} \right) (y - \bar{y}) d\Gamma + \int_{\Omega} [\bar{s} + \bar{p}][A(y - \bar{y})] dx \geq 0,$$

$$(4.10) \quad \forall v \in U_{\text{ad}} \quad \int_{\Gamma} (N\bar{v} - \bar{q})(v - \bar{v}) d\Gamma \geq 0.$$

Proof. As in § 3, q_ε and s_ε are bounded in \mathbb{L}^2 . So we are allowed to pass to the (weak) limit in the penalized optimality system. The rest of the proof is exactly the same. \square

Remark 14. Let us compare the condition (\mathcal{H}) to the Slater condition (\mathcal{S}) for the boundary case and assume for simplicity $f = 0$. It holds that

$$\mathbf{P}(v) \begin{cases} Ay = 0 & \text{in } \Omega, \\ y = v & \text{on } \Gamma, \end{cases}$$

has a unique solution $y = T(v)$, in \mathbb{W} for any v in $\mathbb{H}^{-1/2}(\Gamma)$.

The Slater condition, $0 \in \text{Int}_W(T(U_{\text{ad}}) - K)$, is equivalent to

$$(\mathcal{S}) \quad \exists \rho > 0, \quad \forall \varphi \text{ s.t. } \|\varphi\|_W \leq 1, \quad \exists v_\varphi \in U_{\text{ad}}, \quad \exists y_\varphi \in K, \quad y_\varphi = T(v_\varphi) - \rho\varphi,$$

and (\mathcal{H}) is

$$\begin{aligned} \exists R > 0, \quad \exists \rho > 0, \quad \exists v_o \in U_{\text{ad}}, \\ \forall \psi = (\kappa, \xi) \in \mathcal{B}^2(\Omega) \times \mathcal{B}^2(\Gamma), \quad \exists v_\psi \in \mathcal{B}^2(v_o, R) \cap U_{\text{ad}}, \\ \exists y_\psi \in K \text{ such that} \\ Ay_\psi = 0 \quad \text{in } \Omega, \\ y_\psi = v \quad \text{on } \Gamma. \end{aligned}$$

If U_{ad} is not bounded, (\mathcal{H}) is more restrictive than (\mathcal{S}) because we impose v_ψ in $\mathcal{B}^2(v_o, R)$. If U_{ad} is bounded, $(\mathcal{H}) \Leftrightarrow (\mathcal{S})$.

Proof. $(\mathcal{S}) \Rightarrow (\mathcal{H})$.

Let $\psi = (\kappa, \xi)$ be in $\mathcal{B}^2(\Omega) \times \mathcal{B}^2(\Gamma)$. $\exists! \varphi \in \mathbb{W}$ such that $A\varphi = \kappa$ in Ω and $\varphi = \xi$ on Γ . Moreover, $\|\varphi\|_W \leq C$ (because T is an isomorphism). So there exists $(v_\varphi, y_\varphi) \in U_{\text{ad}} \times K$ such that $y_\varphi = T(v_\varphi) - (\rho/C)\varphi$, and (\mathcal{H}) is satisfied (with ρ/C instead of ρ).

$(\mathcal{H}) \Rightarrow (\mathcal{S})$. Let $\varphi \in \mathbb{W}$ such that $\|\varphi\|_W \leq 1$, $A\varphi \in \mathbb{L}^2(\Omega)$, and $\|A\varphi\|_\Omega \leq 1$. So there exists $(v^*, y^*) \in U_{\text{ad}} \times K$ such that $Ay^* = -\rho A\varphi$ in Ω and $y^* = v^*$ on Γ . Then (\mathcal{S}) is satisfied with $y_\varphi = T(v^*) - \rho\varphi$. \square

Now we may weaken this result as we did in the previous section; we choose $\mathcal{E}_\Omega = \mathcal{C}_o(\Omega)$ and $\mathcal{E}_\Gamma = \mathcal{C}_o(\Gamma)$ (with the L^∞ -norm), so that (4.8) becomes (\mathcal{H}^*) ,

$$\begin{aligned} \exists \rho > 0, \quad \exists v_o \in U_{\text{ad}}, \quad \exists R > 0 \quad \text{such that} \\ \forall \eta = (\kappa, \xi) \in \mathcal{B}^\infty(\mathcal{C}_o(\Omega)) \times \mathcal{B}^\infty(\mathcal{C}_o(\Gamma)), \quad \exists v_\eta \in \mathcal{B}^2(v_o, R) \cap U_{\text{ad}} \quad \text{such that} \\ \text{the solution } y_\eta \text{ of } (P_\eta) \text{ belongs to } K; \\ (\mathcal{B}^\infty(\mathcal{C}_o(\Omega)) = \{\kappa \in \mathcal{C}_o(\Omega) \mid \|\kappa\|_{L^\infty(\Omega)} \leq 1\}). \end{aligned}$$

The difficulties we met in the distributed case still remain, of course. We may bound the multipliers in a measure space and pass to the limit in the dual sense $\langle \text{measures}, \mathcal{C}_o \text{ functions} \rangle$. So we must consider, once again, test functions that are smooth enough. Let us define

$$\begin{aligned} \mathcal{F}(\bar{y}) &= \{y \in \mathbb{W} \mid \exists (\varphi, \psi) \in \mathcal{C}_o(\Omega) \times \mathcal{C}_o(\Gamma), Ay = A\bar{y} + \varphi \text{ on } \Omega, y = \bar{y} + \psi \text{ on } \Gamma\}, \\ \mathcal{G}(\bar{v}) &= \bar{v} + \mathcal{C}_o(\Gamma). \end{aligned}$$

We then have the following theorem.

THEOREM 4.6. *Let (\bar{y}, \bar{v}) be the optimal solution of (Π) and assume (\mathcal{H}^*) . Then there exists $\bar{q} \in \mathcal{M}(\Gamma)$, $\bar{s} \in \mathcal{M}(\Omega)$ such that*

$$(4.11) \quad \forall y \in K^* \cap \mathcal{F}(\bar{y}) \quad \int_\Gamma \left(-\frac{\partial \bar{p}}{\partial \nu_{A^*}} + \bar{q} \right) (y - \bar{y}) \, d\Gamma + \int_\Omega [\bar{s} + \bar{p}][A(y - \bar{y})] \, dx \geq 0,$$

$$(4.12) \quad \forall v \in U_{\text{ad}} \cap \mathcal{G}(\bar{v}) \quad \int_\Gamma (N\bar{v} - \bar{q})(v - \bar{v}) \, d\Gamma \geq 0,$$

where $\mathcal{M}(\Omega)$ (respectively, $\mathcal{M}(\Gamma)$) is the set of bounded Radon measures on Ω (respectively, on Γ).

Remark 15. (\mathcal{H}^*) is weaker than (\mathcal{S}) . Indeed, if we take φ in \mathbb{W} such that $\|\varphi\|_W \leq 1$, we cannot always find v_φ in U_{ad} and y_φ in K (with the assumption (\mathcal{H}^*)) such that $Ay_\varphi = -\rho A\varphi$ in Ω because, generally, $A\varphi$ does not belong to $\mathbb{L}^\infty(\Omega)$. Anyway, the distinction between (\mathcal{H}) (i.e., (\mathcal{S})) and (\mathcal{H}^*) does not seem to be useful (at least in the case of pointwise state constraints) because it is essentially a distinction on interior perturbations, and we have a boundary problem.

Now let us end this remark with the following example:

$$K = \{y \in \mathbb{W} \mid \varphi \leq y \leq \psi \text{ a.e. in } \Omega\}$$

and

$$U_{\text{ad}} = \{v \in \mathbb{L}^2(\Gamma) \mid \alpha \leq v \leq \beta \text{ a.e. on } \Gamma\}$$

(with $\alpha \leq \varphi|_{\Gamma}$ and $\beta \geq \psi|_{\Gamma}$). If we can find $e \in \mathcal{C}^0(\bar{\Omega})$, $e > 0$, such that $\varphi + e \leq \psi - e$, then (\mathcal{H}) is satisfied (with $y_o \in [\varphi + e, \psi - e]$ and $v_o = y_o|_{\Gamma}$).

4.2. Neumann boundary conditions. We now consider the following *Neumann* problem:

$$(4.13) \quad \mathbf{P}(v) \begin{cases} Ay = f & \text{in } \Omega, \\ \frac{\partial y}{\partial v_A} = v & \text{on } \Gamma, \end{cases}$$

where $f \in \mathbb{L}^2(\Omega)$ and $v \in \mathbb{L}^2(\Gamma)$.

The unique solution $y = T(v)$ of $\mathbf{P}(v)$ is in $\mathbb{V} = \{y \in \mathbb{H}^1(\Omega) \mid Ay \in \mathbb{L}^2(\Omega)\}$. K and U_{ad} are nonempty, closed, convex, subsets of, respectively, \mathbb{V} and $\mathbb{L}^2(\Gamma)$. If y belongs to \mathbb{V} , then $(\partial y / \partial v_A)$ belongs to $\mathbb{H}^{-1/2}(\Gamma)$, but not necessarily to $\mathbb{L}^2(\Gamma)$. So we define, as before, the convex subset $K^* = \{y \in K \mid (\partial y) / (\partial v_A) \in \mathbb{L}^2(\Gamma)\}$.

Remark 16. If $U_{\text{ad}} \subseteq \mathbb{H}^{1/2}(\Gamma)$, then $T(v) \in \mathbb{H}^2(\Omega)$ and $K^* = K$.

Let us give $\varepsilon > 0$ and define J_ε on $\mathbb{V} \times \mathbb{L}^2(\Gamma)$ with

$$(4.14) \quad J_\varepsilon(y, v) = \begin{cases} J(y, v) + \frac{1}{2\varepsilon} \int_{\Omega} (Ay - f)^2 dx + \frac{1}{2\varepsilon} \int_{\Gamma} \left(\frac{\partial y}{\partial v_A} - v \right)^2 d\Gamma, & \text{if } \frac{\partial y}{\partial v_A} \in \mathbb{L}^2(\Gamma), \\ +\infty & \text{if not.} \end{cases}$$

PROPOSITION 4.7. *The penalized problem (Π_ε) has a unique optimal solution $(y_\varepsilon, v_\varepsilon)$ and for all $(y, v) \in K^* \times U_{\text{ad}}$ $J'_\varepsilon(y_\varepsilon, v_\varepsilon)(y - y_\varepsilon, v - v_\varepsilon) \geq 0$.*

We define (as usual)

$$(4.15) \quad \left[\begin{array}{l} s_\varepsilon = \frac{Ay_\varepsilon - f}{\varepsilon} \in \mathbb{L}^2(\Omega), \\ q_\varepsilon = \frac{1}{\varepsilon} \left(\frac{\partial y_\varepsilon}{\partial v_A} - v_\varepsilon \right) \in \mathbb{L}^2(\Gamma), \\ \text{and } p_\varepsilon \in \mathbb{H}^2(\Omega) \text{ (adjoint state), solution of} \\ A^* p_\varepsilon = y_\varepsilon - z_d \text{ in } \Omega, \\ \frac{\partial p_\varepsilon}{\partial v_{A^*}} = 0 \text{ on } \Gamma, \end{array} \right.$$

and we still obtain results similar to previous ones.

PROPOSITION 4.8. *Let $(y_\varepsilon, v_\varepsilon)$ be the optimal solution of (Π_ε) :*

$$(4.16) \quad \forall y \in K^* \quad \int_{\Gamma} (p_\varepsilon + q) \frac{\partial(y - y_\varepsilon)}{\partial v_A} d\Gamma + \int_{\Omega} [s_\varepsilon + p_\varepsilon][A(y - y_\varepsilon)] dx \geq 0,$$

$$(4.17) \quad \forall v \in U_{\text{ad}} \quad \int_{\Gamma} (Nv_\varepsilon - q_\varepsilon)(v - v_\varepsilon) d\Gamma \geq 0.$$

THEOREM 4.9. As ε tends to 0,

- (i) v_ε converges to \bar{v} strongly in $\mathbb{L}^2(\Gamma)$;
- (ii) y_ε to \bar{y} strongly in \mathbb{V} (where (\bar{y}, \bar{v}) is the solution of (Π));
- (iii) p_ε converges to \bar{p} strongly in $\mathbb{H}^2(\Omega)$, where \bar{p} is solution of

$$(4.18) \quad \begin{aligned} A^* \bar{p} &= \bar{y} - z_d \quad \text{in } \Omega, \\ \frac{\partial \bar{p}}{\partial v_{A^*}} &= 0 \quad \text{on } \Gamma. \end{aligned}$$

Once again, we must estimate q_ε and s_ε , and we assume that

$$(4.19) \quad \begin{aligned} &\exists \rho > 0, \quad \exists v_o \in U_{\text{ad}}, \quad \exists R > 0 \quad \text{such that} \\ &\forall \eta = (\kappa, \xi) \in \mathcal{B}(\mathcal{E}_\Omega) \times \mathcal{B}(\mathcal{E}_\Gamma), \quad \exists v_\eta \in \mathcal{B}^2(v_o, R) \cap U_{\text{ad}} \quad \text{such that } y_\eta \in K, \end{aligned}$$

where y_η is the solution of

$$(P_\eta) \quad \begin{cases} Ay_\eta = f - \rho\kappa & \text{in } \Omega, \\ \frac{\partial y_\eta}{\partial v_A} = v_\eta - \rho\xi & \text{on } \Gamma. \end{cases}$$

4.2.1. The “strong” optimality system.

THEOREM 4.9. Let us assume (4.19) with $\mathcal{E}_\Omega = \mathbb{L}^2(\Omega)$ and $\mathcal{E}_\Gamma = \mathbb{L}^2(\Gamma)$ and let (\bar{y}, \bar{v}) be in \mathcal{D} . (\bar{y}, \bar{v}) is the optimal solution of (Π) if and only if there exists $\bar{q} \in \mathbb{L}^2(\Gamma)$, $\bar{s} \in \mathbb{L}^2(\Omega)$ such that

$$(4.20) \quad \forall y \in K^* \quad \int_\Gamma (\bar{p} + \bar{q}) \frac{\partial(y - \bar{y})}{\partial v_A} d\Gamma + \int_\Omega (\bar{s} + \bar{p})[A(y - \bar{y})] dx \geq 0,$$

$$(4.21) \quad \forall v \in U_{\text{ad}} \quad \int_\Gamma (N\bar{v} - \bar{q})(v - \bar{v}) d\Gamma \geq 0,$$

where \bar{p} is the adjoint state defined by (4.18).

4.2.2. The “weak” optimality system. In this case, $\mathcal{F}(\bar{y})$ and $\mathcal{G}(\bar{v})$ are defined as follows:

$$\mathcal{F}(\bar{y}) = \left\{ y \in \mathbb{V} \mid \exists (\varphi, \psi) \in \mathcal{C}_o(\Omega) \times \mathcal{C}_o(\Gamma), Ay = A\bar{y} + \varphi \text{ on } \Omega, \frac{\partial y}{\partial v_A} = \frac{\partial \bar{y}}{\partial v_A} + \psi \text{ on } \Gamma \right\},$$

$$\mathcal{G}(\bar{v}) = \bar{v} + \mathcal{C}_o(\Gamma).$$

THEOREM 4.10. Let us assume (4.19) with $\mathcal{E}_\Omega = \mathcal{C}_o(\Omega)$ and $\mathcal{E}_\Gamma = \mathcal{C}_o(\Gamma)$ (with the \mathbb{L}^∞ -norm, so that $\mathcal{B}(\mathcal{E}) = \{\kappa \in \mathcal{C}_o \mid \|\kappa\|_{\mathbb{L}^\infty} \leq 1\}$). Let (\bar{y}, \bar{v}) be the optimal solution of (Π) ; then there exists $\bar{q} \in \mathcal{M}(\Gamma)$, $\bar{s} \in \mathcal{M}(\Omega)$ such that

$$(4.22) \quad \forall y \in K^* \cap \mathcal{F}(\bar{y}) \quad \int_\Gamma (\bar{p} + \bar{q}) \frac{\partial(y - \bar{y})}{\partial v_A} d\Gamma + \int_\Omega (\bar{s} + \bar{p})[A(y - \bar{y})] dx \geq 0,$$

$$(4.23) \quad \forall v \in U_{\text{ad}} \cap \mathcal{G}(\bar{v}) \quad \int_\Gamma (N\bar{v} - \bar{q})(v - \bar{v}) d\Gamma \geq 0,$$

where \bar{p} is the adjoint state defined by (4.18).

5. Conclusion. For every “classical” case we sought, we retrieve already-known results (i.e., the existence of an optimality system, assuming the Slater condition). In the case where the control is distributed, we have proved the existence of multipliers that are measures (if the Slater condition is not satisfied). The main problem then is to “estimate” the sets $\mathcal{F}(\bar{y})$ and $\mathcal{G}(\bar{v})$ defined for the “weak” system: are they “large

enough" to allow good numerical experimentation? The answer is not obvious; we must study how Lagrangian methods (for example) work and the regularity of the solution (\bar{y}, \bar{v}) .

This penalization method may be applied to any problem of control with state constraints (especially to nonlinear problems) and provides decoupled optimality systems that may be solved with classical multipliers methods. Moreover, the penalization itself may lead to algorithms that allow us to compute the solution and the multiplier simultaneously. Once again, the regularity of the multiplier is one of the main difficulties we are meeting, and the optimality system may answer this kind of question (see [13] and [2], for instance).

REFERENCES

- [1] F. ABERGEL, *Conditions d'optimalité pour un problème de contrôle mal posé*, Compte Rendu de l'Académie des Sciences, t. 303, Sér. I, nos. 6 et 7, 1986, pp. 295–301.
- [2] F. ABERGEL AND R. TEMAM, *Optimality condition for some nonqualified problems of distributed control*, SIAM J. Control Optim., 1 (1989), pp. 1–12.
- [3] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
- [4] V. BARBU, *Boundary control problems with convex cost criterion*, SIAM J. Control Optim., 2 (1980), pp. 227–283.
- [5] M. BERGOUNIOUX, *Etude de différents problèmes de contrôle avec contraintes sur l'état*, Rapport de Recherche 89-1, Université d'Orléans, France, 1989.
- [6] ———, *Problèmes de contrôle avec contraintes sur l'état*, Compte Rendu de l'Académie des Sciences, t. 310, Série I, 1990, pp. 391–396.
- [7] J. F. BONNANS AND E. CASAS, *Contrôle de systèmes non linéaires comportant des contraintes distribuées sur l'état*, Rapport de Recherche 300, INRIA, Le Chesnay, France, 1984.
- [8] ———, *On the choice of the function space for some state constrained control problems*, Numer. Funct. Anal. Optim., 4 (1984–1985), pp. 333–348.
- [9] ———, *Quelques méthodes pour le contrôle optimal de problèmes comportant des contraintes sur l'état*, An. Ştiinţ. Univ. "Al. I. Cuza" Iaşi Sect. I a Mat., 32, 3 (1986), pp. 57–62.
- [10] ———, *Optimal control of semilinear multistate systems with state constraints*, SIAM J. Control Optim., 2 (1989), pp. 446–455.
- [11] H. BREZIS, *Analyse Fonctionnelle. Théorie et Applications*, Masson, Paris, 1983.
- [12] E. CASAS, *Quelques problèmes de contrôle avec contraintes sur l'état*, Compte Rendu de l'Académie des Sciences, 296, Sér. I, 1983, p. 509.
- [13] ———, *Control of elliptic problem with pointwise state constraints*, SIAM J. Control Optim., 6 (1986), pp. 1309–1322.
- [14] I. EKELAND AND R. TEMAM, *Analyse convexe et problèmes variationnels*, Dunod, Gauthier-Villars, Paris, 1974.
- [15] M. FORTIN AND R. GLOWINSKI, *Méthodes de Lagrangian augmenté*, Collection Méthodes Mathématiques pour l'Informatique, Dunod, Paris, 1982.
- [16] I. LASIECKA, *State constrained control problems for parabolic systems: Regularity of optimal solutions*, Appl. Math. Optim., 6 (1980), pp. 1–29.
- [17] J. L. LIONS, *Contrôle Optimal des Systèmes Gouvernés par des Equations aux Dérivées Partielles*, Dunod, Gauthier-Villars, Paris, 1968.
- [18] ———, *Contrôle des systèmes distribués singuliers*, Gauthier-Villars, Paris, 1983.
- [19] J. L. LIONS AND E. MAGENES, *Problèmes aux Limites Non Homogènes et Applications*, Dunod, Gauthier-Villars, Paris, 1968.
- [20] U. MACKENROTH, *Convex parabolic boundary control problems with pointwise state constraints*, J. Math. Anal. Appl., 87 (1982), pp. 256–277.
- [21] ———, *On some elliptic optimal control problems with state constraints*, Optim., 5 (1986), pp. 595–607.
- [22] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [23] L. WHITE, *Control of a hyperbolic problem with pointwise state constraints*, J. Optim. Theory Appl., 2 (1983), pp. 359–369.

DIFFERENTIAL INCLUSIONS AND TARGET PROBLEMS*

MARC QUINCAMPOIX†

Abstract. Where and how solutions associated to a differential inclusion can or cannot enter a given target is studied. For this purpose, partitions of the target boundary are associated with the dynamic of the system. The behaviour of these solutions is qualitatively described in terms of viability and invariance kernels of sets. These kernels determine points such that there exist (respectively, all) solutions starting at these points remain in a given set of constraints. The sets that are reached in finite time by viable solutions to the system are also studied. Finally, some applications to control systems with one target are provided, and the concept of semipermeable barriers will be generalized.

Résumé. Il s'agit de savoir où et comment les solutions d'une inclusion différentielle peuvent ou non franchir le bord d'une cible donnée. On établit pour cela une partition de la frontière de la cible en fonction de la dynamique du système. Ceci sera utilisé pour décrire le comportement qualitatif des trajectoires par rapport aux noyaux de viabilité et d'invariance de plusieurs ensembles. Ces noyaux décrivent les conditions initiales à partir desquelles il existe des solutions de l'inclusion différentielle avec contraintes. Nous définirons aussi des ensembles qui sont atteints en temps fini par des solutions de l'inclusion différentielle avec contraintes. Cet article se termine par une application aux systèmes contrôlés à une cible et par une généralisation de la notion de barrière semi perméable.

Key words. viability, invariance, semipermeable barriers, differential inclusions

AMS(MOS) subject classification. 49A50

1. Introduction. We consider an open set C and a system whose evolution is described by the following differential inclusion:¹

$$x'(t) \in F(x(t)),$$

where F is a set-valued map and $x(t) \in X := \mathbb{R}^n$. We assume throughout this paper, that the set valued map F has nonempty values.²

A first question is: On what part of the boundary can the state of the system reach the target C ?

For that purpose, we prove a new result about the tangent cone to an intersection. Thanks to this, we define the three sets C^e , C^b , C^i of the boundary of C that form a partition of ∂C . Then we prove that solutions crossing the three areas C^e , C^b , C^i of ∂C have the following qualitative properties:

- If a solution crosses C^i , then it enters C ,
- If a solution crosses C^e , then it goes outside C ,
- If a solution crosses the interior³ of C^b , then it remains (locally) in ∂C .

In answering the above question, we prove a new result about the tangent cone to an intersection. Thanks to this, the three sets C^e , C^b , C^i form a partition of the boundary of C .

* Received by the editors August 29, 1990; accepted for publication (in revised form) March 1, 1991.

† Centre de la recherche de mathématique de la décision (CEREMADE), Université Paris-Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 Paris cedex 16.

¹ This allows us to incorporate a lack of exact knowledge of dynamics or to represent a control system with state-dependent control map $U(x)$:

$$\begin{aligned} x'(t) &= f(x(t), u(t)), \\ u(t) &\in U(x(t)) \end{aligned}$$

through a differential inclusion, by setting: $F(x) := f(x, U(x))$.

² If it is not the case, we can study the differential inclusion in the interior of the domain of F .

³ In the relative topology of ∂C .

Another question is: From what initial conditions can the system reach C ?
Let us consider a differential inclusion with constraints

$$(1) \quad x'(t) \in F(x(t)), \quad x(t) \in K,$$

where K is a given closed set (we shall use $K := X \setminus C$).

We need some definitions and properties concerning differential inclusions with constraints (see [5], [6]).

We shall say that a solution $x(\cdot)$ of the differential inclusion (1) is *viable* in K if and only if

$$\forall t \geq 0, \quad x(t) \in K.$$

The solution $x(\cdot)$ is *locally viable* in K if and only if

$$\exists T > 0, \quad \text{such that } \forall t \leq T, \quad x(t) \in K.$$

A set K has the *viability property* if and only if, for any point x_0 of K , there exists at least one solution to (1) starting at this point that is viable in K .

A set K has the *invariance property* if and only if for any point x_0 of K , all solutions to (1) starting at this point are viable in K .

In the literature, the above properties appeared in different contexts under many names (when F is single-valued the both reduce to invariance of dynamical systems). For general control systems introduced by Roxin, these two properties have different names. “Viability property” is called “weak invariance,” and “invariance property” is called “strong invariance” (see [26], [27]). They also have been called “controlled invariance” by other authors (Wonham, Byrnes, Isidori, Morse, and others).

A closed set K is a *viability domain* if and only if

$$\forall x \in K, \quad F(x) \cap T_K(x) \neq \emptyset.$$

The set K is an *invariance domain* if and only if

$$\forall x \in K, \quad F(x) \subset T_K(x),$$

where $T_K(x)$ denotes the contingent cone⁴ to K at x .

If K is closed, if F is an upper semicontinuous⁵ set-valued map with nonempty closed convex compact values and linear growth,⁶ then, thanks to Haddad’s viability theorem (see [17], [5]), K is a viability domain if and only if the viability property holds for K .

When K is not a viability domain, the question arises as to how to find closed subsets in which it is possible to solve (1) in K . With these assumptions it is possible to define the *viability kernel*.

DEFINITION 1.1. The viability kernel of a closed set K is the largest closed viability domain contained in K .

We have some examples of computation of viability kernels in [7] and [14].

⁴ Recall that $T_K(x) := \{v \in X \mid \liminf_{h \rightarrow 0^+} d(x + hv, K)/h = 0\}$.

⁵ Let us recall that a set-valued map F is upper semicontinuous at x_0 if and only if

$$\forall \varepsilon > 0, \exists \alpha > 0, \quad F(x_0 + \alpha B) \subset F(x_0) + \varepsilon B.$$

⁶ We say that a map F has a linear growth if there exists $c > 0$ such that

$$\forall x \in X, \quad F(x) \subset c(1 + \|x\|)B.$$

In a similar way, thanks to the invariance theorem (see [5]), if K is closed and if F is Lipschitzian⁷ with compact convex values and linear growth, then K is an invariance domain if and only if the invariance property holds for K .

Under these assumptions it is possible to define (see [5]) the *invariance kernel*.

DEFINITION 1.2. The invariance kernel of a closed set K is the largest closed invariance domain contained in K .

In this paper, we prove some properties of the boundaries of viability and invariance kernels. In fact, under adequate assumptions, the boundaries of $\text{Viab}_F(K)$ and $\text{Inv}_F(K)$ are viability domains.

We shall also try to answer another question: Is it possible to find a solution to the differential inclusion with constraints that reaches a given point? To investigate this question, we shall introduce and study kernels for $-F$ that yields backward trajectories.

In the last section, results concerning the boundaries of the kernels of a differential inclusion will be used to study the following control system with a target C :

$$x'(t) = f(x(t), u(t)) \quad u(t) \in U(x(t)).$$

We shall generalize the concept of *semipermeable barrier* (introduced by Isaacs in [18] for differential games). Recall that a barrier allows us to separate the areas from which it is possible to reach C and the areas from which it is not possible (see also [10], [9], [11]). Recall that a C^1 -surface is semipermeable when it satisfies an equation such that $\max_u f(x, u) \cdot n \leq 0$ or $\min_u f(x, u) \cdot n \geq 0$ (where n is the normal vector of the surface). It means that the solutions of § 5.1, (5) are able to cross the surface in “only one direction.” In fact, we prove that the solutions of a control system can cross the boundaries of viability and invariance kernels only from the exterior of the kernel to the interior of the kernel. *In this sense* the boundaries of invariance and viability kernels are semipermeable.

2. The target boundary and the dynamics. We study a system whose dynamics are described by the differential inclusion:

$$(2) \quad x'(t) \in F(x(t)),$$

where F is the set-valued map whose values are nonempty convex and compact from a finite-dimensional vector space X into itself. We also consider a set C (the target) that is open nonempty and different from X .

Let us define two closed sets $K := X \setminus C$ and $\hat{K} := \overline{X \setminus K} = \bar{C}$. The Haddad viability theorem [17] provides conditions such that the state never reaches C . Here we study how it is possible to reach C .

We first state our results; their proofs will be given in § 2.4.

2.1. A geometrical result. We need a result concerning the contingent cone to an intersection of two closed sets.

DEFINITION 2.1. Let K be a closed set. The Dubovitsky–Milliutin tangent cone is defined by

$$D_K(x) := \{v \in X \mid \exists \alpha > 0, x +]0, \alpha][v + \alpha B) \subset K\}$$

or, equivalently, $D_K(x) = X \setminus T_{\overline{X \setminus K}}(x)$.

⁷ Let us recall (see [5], [6]) that a set-valued F is Lipschitzian if and only if there exists a positive real k such that

$$\forall (x, y) \in X \times X, \quad F(x) \subset F(y) + k\|x - y\|B.$$

THEOREM 2.2. *Let K_1 and K_2 be two closed sets of a normed vector space X . Then, for any x ,*

$$T_{K_1}(x) \cap T_{K_2}(x) \cap D_{K_1 \cup K_2}(x) \subset T_{K_1 \cap K_2}(x).$$

This result allows us to characterize the intersection of contingent cones, without assumptions on the regularity of these cones.⁸

COROLLARY 2.3. *Let K_1 and K_2 be two closed subsets of X .*

If $x \in K_1 \cap K_2 \cap \text{Int}(K_1 \cup K_2)$, then

$$T_{K_1}(x) \cap T_{K_2}(x) = T_{K_1 \cap K_2}(x).$$

We recall that the same conclusion can be obtained when we assume the following transversality condition:

$$(3) \quad C_{K_1}(x) - C_{K_2}(x) = X,$$

where $C_K(x)$ denotes the Clarke's cone⁹ to K at x .

Remark. These two results allow us to express the contingent cone to an intersection in different cases.

In \mathbb{R}^2 , we can compute the tangent cone at $(0, 0)$ to the intersection of

$$K_1 := \{(x, y) \mid x \leq 0\} \text{ and } K_2 := \{(x, y) \mid x \geq 0\}$$

thanks to Corollary 2.3, but not from the transversality condition (3). In the case of

$$K_1 := \{(x, y) \mid x = y\} \text{ and } K_2 := \{(x, y) \mid x = -y\},$$

it is possible to compute the intersection thanks to (3), but we cannot use Corollary 2.3.

We can deduce the very useful Corollary 2.4.

COROLLARY 2.4. *Let x belong to X . It holds that*

$$T_K(x) \cap T_{\hat{K}}(x) = T_{\partial K}(x)$$

and

$$D_K(x) = T_K(x) \setminus T_{\partial K}(x).$$

This corollary can be used in the study of the qualitative behaviour of replicator systems in the simplex (see [12]), or to study the fluctuations of solutions around the boundary of a given set (see [19]). Corollary 2.3 can be generalized to compute the contingent cone to an intersection of a finite number of closed sets as follows.

COROLLARY 2.5. *Let K_1, K_2, \dots, K_p be p closed subsets of a metric vector space X and let x belong to $\bigcap_{i=1}^p K_i$. If there exists an open set \mathcal{O} that contains x such that*

$$\forall j \leq p, \quad \mathcal{O} \subset K_j \cup \left(\bigcap_{i=1}^{j-1} K_i \right),$$

then

$$T_{\bigcap_{i=1}^p K_i}(x) = \bigcap_{i=1}^p T_{K_i}(x).$$

⁸ It is a pleasure for me to thank Halina Frankowska who suggested improving Corollary 2.3 into Theorem 2.2 by using Dubovitsky-Millutin tangent cones.

⁹ Recall the definition of the Clarke's tangent cone (see, for instance, [6, Chap. 4]), $C_K(x) = \{v \mid \liminf_{h \rightarrow 0^+, y \rightarrow x, y \in K} d(y \geq hv, K)/h = 0\}$.

2.2. First partition of the target boundary. Let us introduce three subsets of the boundary that are dependent on the dynamic of the system

$$\begin{aligned} K^i &:= \{x \in \partial K / F(x) \subset D_K(x)\}, \\ K^e &:= \{x \in \partial K / F(x) \subset X \setminus T_K(x)\}, \\ K^b &:= \{x \in \partial K / F(x) \cap T_{\partial K}(x) \neq \emptyset\}. \end{aligned}$$

PROPOSITION 2.6. *If $F: X \rightarrow X$ is an upper semicontinuous set-valued map with nonempty convex compact values, if K is closed nonempty, then (K^e, K^i, K^b) form a partition of the boundary ∂K (in the sense that $\partial K = K^i \cup K^b \cup K^e$ and these three sets are disjoint).*

If x_0 belongs to K^i , then all solutions starting at x_0 enter $\text{Int}(K)$ and stay in the interior on time interval $]0, T[$ (with $T > 0$).

If x_0 belongs to K^e , then all solutions starting at x_0 enter $\text{Int}(X \setminus K)$ and stay outside K on $]0, T[$ (with $T > 0$).

If x_0 belongs¹⁰ to $\text{Int}_{\partial K} K^b$, then there exists a solution starting at x_0 that stays on the boundary ∂K on $]0, T[$ (with $T > 0$).

Remark. We can note that, when ∂K is a C^1 surface, the subset $K^b \cup K^i$ is often called the *boundary usable part* and K^e the *boundary nonusable part*.

For a set A , we set $\hat{A} := \overline{X \setminus A}$ and we denote by $\text{Int}(A)$ its interior. We can introduce the same type of partition for the closed set $\bar{C} = \hat{K}$, i.e., the three sets $\bar{C}^i = \hat{K}^i$, $\bar{C}^b = \hat{K}^b$, $\bar{C}^e = \hat{K}^e$, which form a partition of $\partial \hat{K}$. A natural question arises: How can we compare \hat{K}^i , \hat{K}^b , \hat{K}^e and (K^e, K^i, K^b) ?

PROPOSITION 2.7. *Let K be a closed set and F a set-valued map with nonempty convex values.*

$$K^i = \hat{K}^e, K^e \subset \hat{K}^i, K^b \subset \hat{K}^b.$$

Equalities hold true if and only if $\overline{\text{Int}(\bar{K})} = K$.

The first statement and the inclusions are obvious. If the equalities $K^e = \hat{K}^i$, $K^b = \hat{K}^b$ hold, then necessarily $\partial K = \partial \hat{K}$, i.e., $\overline{\text{Int}(\bar{K})} = K$. It is easy to show that this condition is sufficient (in this case $\hat{K} = K$).

We can improve this partition to have a more precise one.

2.3. Second partition of the target boundary. Let us consider the following differential inclusion:

$$(4) \quad y'(t) \in -F(y(t)).$$

We can regard the solutions of (4) as solutions of (1) but in the reverse direction (i.e., if $x(\cdot)$ is a solution of (1) on $[0, T]$, then $y(t) := x(T - t)$ is solution of (4) on $[0, T]$). In this way, we get the backward trajectories of (1).

We introduce the subsets

$$\begin{aligned} K^{i-} &:= \{x \in \partial K / -F(x) \subset D_K(x)\}, \\ K^{e-} &:= \{x \in \partial K / -F(x) \subset X \setminus T_K(x)\}, \\ K^{b-} &:= \{x \in \partial K / -F(x) \cap T_{\partial K}(x) \neq \emptyset\}. \end{aligned}$$

These three sets also form a partition of ∂K . Consequently, these two partitions yield a new partition of the boundary made of nine subsets.

¹⁰ Here, we denote by $\text{Int}_{\partial K} K^b$ the interior of K^b in the space ∂K .

We can describe the qualitative behaviour of solution, as in the previous section, in the following proposition in which we shall denote by $\text{Int}(K^b)$ the interior of K^b in the space ∂K .

PROPOSITION 2.8. *Let K be a closed nonempty set.*

x_0 is an element of	Properties of solutions that start at x_0
$K^i \cap K^{i-}$	All solutions enter K at x_0 .
$K^i \cap K^{e-}$	No solutions come from the exterior of K . All solutions enter K at x_0 .
$K^i \cap \text{Int}(K^{b-})$	No solutions come from the interior of K . All solutions enter K at x_0 .
$K^e \cap K^{e-}$	There exists at least one trajectory locally viable on the boundary that comes into $\text{Int}(K)$ at x_0 . All solutions go outside K .
$K^e \cap K^{i-}$	No solutions come from the interior of K . All solutions go outside K .
$K^e \cap \text{Int}(K^{b-})$	No solutions come from the exterior of K . All solutions go outside K .
$\text{Int}(K^b) \cap \text{Int}(K^{b-})$	There exists at least one solution locally viable on the boundary that comes into $\text{Int}(K)$ at x_0 . There exists a solution passing through x_0 (i.e., $\exists \tau > 0$, $x(\tau) = x_0$) and locally viable (for F and $-F$) on the boundary.
$\text{Int}(K^b) \cap K^{i-}$	No solutions come from the exterior of K .
$\text{Int}(K^b) \cap K^{e-}$	There exists a solution locally viable on ∂K that comes from the interior of K . No solutions come from the interior of K . There exists a solution locally viable on ∂K that comes from the exterior of K .

2.4. Proofs.

Proof of Theorem 2.2. Let v be in $T_{K_1}(x) \cap T_{K_2}(x) \cap D_{K_1 \cup K_2}(x)$. According to the definitions of these sets, there exist sequences h_n^1, h_n^2 of nonnegative reals converging to 0, sequences v_n^1, v_n^2 converging to v , and a real α such that

$$\forall n \quad x + h_n^1 v_n^1 \in K_1, \quad x + h_n^2 v_n^2 \in K_2, \\ x +]0, \alpha](v + \alpha B) \subset K_1 \cup K_2.$$

Clearly, there exists N such that

$$\forall n > N, \quad x + h_n^i v_n^i \in x + [0, \alpha](v + \alpha B) \quad \text{for } i = 1, 2.$$

Since for all $n > N$, the two points $x + h_n^1 v_n^1$ and $x + h_n^2 v_n^2$ belong to the convex set $x + [0, \alpha](x + \alpha B)$

$$[x + h_n^1 v_n^1, x + h_n^2 v_n^2] \subset x + [0, \alpha](x + \alpha B) \subset K_1 \cup K_2.$$

On the other hand,

$$([x + h_n^1 v_n^1, x + h_n^2 v_n^2] \cap K_1) \cup ([x + h_n^1 v_n^1, x + h_n^2 v_n^2] \cap K_2) = [x + h_n^1 v_n^1, x + h_n^2 v_n^2].$$

We cannot form a partition of a connected set made of two nonempty closed sets. Hence the intersection of the two closed sets in the left-hand side of the above equality is nonempty. Consequently, there exists λ_n in $[0, 1]$ such that

$$\lambda_n(x + h_n^1 v_n^1) + (1 - \lambda_n)(x + h_n^2 v_n^2) \in K_1 \cap K_2$$

(if one of these sets is empty, we obtain the same conclusion by setting $\lambda_n := 0$ or 1).

By setting

$$\begin{aligned} h_n &:= \lambda_n h_n^1 + (1 - \lambda_n) h_n^2, \\ v_n &:= (\lambda_n h_n^1 v_n^1 + (1 - \lambda_n) h_n^2 v_n^2) / h_n, \end{aligned}$$

we see that $v_n \rightarrow v$, $h_n \rightarrow 0$ and $x_n + h_n v_n \in K_1 \cap K_2$. The proof is completed. \square

Corollaries 2.3 and 2.4 are obvious consequences of Theorem 2.2 if we note that

- $D_{K_1 \cup K_2}(x) = X$ by assumption of Corollary 2.3,
- $T_{K_1 \cap K_2}(x) \subset T_{K_1}(x) \cap T_{K_2}(x)$,
- $D_{K \cup \hat{K}}(x) = X$ trivially.

Proof of Corollary 2.5. Thanks to Corollary 2.3, we have

$$T_{K_p}(x) \cap T_{\bigcap_{i=1}^{p-1} K_i}(x) = T_{\bigcap_{i=1}^p K_i}(x).$$

But we have (with the assumption in the case where $j = p - 1$)

$$\emptyset \subset (K_{p-1}) \cup \left(\bigcap_{i=1}^{i=p-2} K_i \right).$$

Hence, according to Corollary 2.3,

$$T_{K_{p-1}}(x) \cap T_{\bigcap_{i=1}^{i=p-2} K_i}(x) = T_{\bigcap_{i=1}^{p-1} K_i}(x).$$

An obvious induction argument allows us to complete the proof. \square

Proof of Proposition 2.6. Thanks to Corollary 2.4, we can divide the space X into three sets: $D_K(x)$, $T_{\partial K}(x)$, $X \setminus T_K(x)$. That provides the partition of the boundary. In fact, we observe that $K^i \cap K^e = \emptyset$ and if $F(x) \cap T_K(x) \neq \emptyset$ and $F(x) \cap T_{\hat{K}}(x) \neq \emptyset$, then using that values of F are connected, we deduce that $x \in K^b$. Now, it is easy to characterize each area (see [5]). \square

3. Boundaries of invariant and viability kernels. In this section, X denotes a finite-dimensional vector space. Our purpose is to describe the boundary of the set of initial conditions of (1) from which it is possible to reach the target C . This set is the complement of the viability kernel of $K = X \setminus C$ associated with (1). We shall now characterize the boundary of these two kernels.

THEOREM 3.1. *Let $F: X \rightarrow X$ be a Lipchitzean set-valued map with nonempty convex compact values and with linear growth, and K be a closed nonempty set.*

If x_0 belongs to $\partial \text{Viab}_F(K) \setminus \partial K$, then there exists a solution viable in K , starting at x_0 that stays in the boundary of $\text{Viab}_F(K)$ as long as it does not cross ∂K . Furthermore, every viable solution starting at x_0 has the same behaviour.

Proof. We prove that there exists a viable solution starting at x_0 that stays on the boundary of the viability kernel until it reaches ∂K .

In fact, let $x(\cdot)$ be a viable solution starting at x_0 that enters the interior of $\text{Viab}_F(K)$ (i.e., there exists $T > 0$ such that $x(T) \in \text{Int}(\text{Viab}_F(K))$ and $x([0, T]) \subset \text{Int}(K) \cap \text{Viab}_F(K)$). According to Filippov's theorem,¹¹ there exists $l > 0$ such that, for all y in K , there exists a solution $y(\cdot)$ starting at y_0 such that

$$\forall t \leq T, \quad y(t) \in x(t) + l \|y_0 - x_0\| B.$$

Hence, it is possible to find $\alpha > 0$ such that, for all y_0 in $(x_0 + \alpha B) \setminus \text{Viab}_F(K)$, we have $y([0, T]) \subset K$ and $y(T) \in \text{Viab}_F(K)$. But there exists a viable solution $\tilde{y}(\cdot)$ starting at $y(T)$ ($\in \text{Viab}_F(K)$).

Let us define the new trajectory $\bar{y}(\cdot)$

$$\bar{y}(s) := \begin{cases} y(s) & \text{if } s \leq T, \\ \tilde{y}(s) & \text{if } s \geq T. \end{cases}$$

¹¹ If F is Lipchitzean with nonempty values, $y(\cdot) \in S_T(y_0)$ (set of solutions of (1) starting at y_0). There exists $l > 0$ such that $d_{W^{1,1}(0,T)}(y, S_T(x_0)) \leq l \|x_0 - y_0\|$ (see [6, Chap. 10]).

Then $\bar{y}(\cdot)$ is a solution to (1) viable in K . We have shown that $\text{Viab}_F(K) \cup \{\bar{y}(t), t \geq 0\}$ (which contains strictly $\text{Viab}_F(K)$) is a viability domain; this is a contradiction. \square

COROLLARY 3.2. *If assumptions of Theorem 3.1 hold true and if $\text{Viab}_F(K) \subset \text{Int}(K)$, then $\partial \text{Viab}_F(K)$ is a viability domain and the set $\overline{X \setminus \text{Viab}_F(K)}$ is an invariance domain.*

Proof. If we note that $\partial \text{Viab}_F(K) \cap \partial K = \emptyset$ then, thanks to Theorem 3.1, $\partial \text{Viab}_F(K)$ is a viability domain. Let us consider a solution $x(\cdot)$ starting at $x_0 \in \overline{X \setminus \text{Viab}_F(K)}$. If a solution reaches $\partial \text{Viab}_F(K)$, thanks to Theorem 3.1, it cannot enter in the interior of the viability kernel. \square

Remark and example. If $\text{Viab}_F(K) \subset \text{Int}(K)$, the sets $\partial \text{Viab}_F(K)$ and $\partial \text{Inv}_F(K)$ are viability and domains, but, generally, they are not invariance domains. We can see that in the following simple examples in the two-dimensional space \mathbb{R}^2 .

We consider a constant set-valued map and two closed sets

$$\begin{aligned} F &:= \{1\} \times [0, 1], \\ K_1 &:= \{(x, y) \in \mathbb{R}^2 \mid x > 0, y \leq 1/x\} \cup \mathbb{R}^- \times \mathbb{R}, \\ K_2 &:= \{(x, y) \in \mathbb{R}^2 \mid x > 0, y \geq -1/x\} \cup \mathbb{R}^- \times \mathbb{R}. \end{aligned}$$

Then, it is easy to check that

$$\text{Viab}_F(K_1) = \mathbb{R} \times \mathbb{R}^- \quad \text{Inv}_F(K_1) = \emptyset,$$

$$\text{Inv}_F(K_2) = \mathbb{R} \times \mathbb{R}^+,$$

$$\text{Viab}_F(K_2) = \{(x, y) \in \mathbb{R}^2 \mid x \geq 1, y \geq -1/x\} \cup \{(x, y) \in \mathbb{R}^2 \mid x \leq 1, y \geq x - 2\}.$$

Here, the boundaries of $\text{Viab}_F(K_1)$ and $\text{Inv}_F(K_2)$ are viability domains, but they are not invariance domains. We can see that all solutions starting at a point of $\{(x, y) \mid x \leq 1, y = x - 2\}$ stay in this set until they reach the boundary of $\text{Viab}_F(K_2)$ (at $(+1, -1)$). \square

We now prove a dual result.

PROPOSITION 3.3. *Let F be a Lipchitzian set-valued map with nonempty convex compact values, and K a closed compact nonempty set.*

If $\text{Inv}_F(K) \subset \text{Int}(K)$, then the boundary $\partial \text{Inv}_F(K)$ is a viability domain.

Proof of Proposition 3.3. We shall show a more precise result: $\overline{X \setminus \text{Inv}_F(K)}$ is a viability domain (so that, if $x_0 \in \partial \text{Inv}_F(K)$ there will exist a solution viable in $\overline{X \setminus \text{Inv}_F(K)}$ that necessarily is also viable $\text{Inv}_F(K)$; hence it is viable in the boundary). In doing this, it is sufficient to show that there exists a solution starting at any point of $\overline{X \setminus \text{Inv}_F(K)}$ that never crosses $\text{Inv}_F(K)$, so that $\overline{X \setminus \text{Inv}_F(K)} \subset \text{Viab}_F(\overline{X \setminus \text{Inv}_F(K)})$.

Since K is a compact set, there exists a nonnegative number α such that $\text{Inv}_F(K) + 2\alpha B \subset \text{Int}(K)$. We shall need the following lemma.

LEMMA 3.4. *Under the assumptions of Proposition 3.3, let $x(\cdot)$ be a solution to (1) and $\alpha > 0$. Then there exists $\tau > 0$ such that, for all $t \geq t' \geq 0$,*

$$x[t', t] \subset K, \quad \|x(t) - x(t')\| \geq \alpha \Rightarrow t - t' \geq \tau.$$

Proof of Lemma 3.4. Let us define $M := \sup_{x \in K} \|F(x)\|$ and $\tau := \alpha/M$. As K is compact and F Lipschitzian with compact values, M is different from infinity.¹² Since

¹² This is even true if we assume that F is upper semicontinuous with compact values.

$x(\cdot)$ is absolutely continuous, we have

$$\alpha \leq \|x(t) - x(t')\| = \left\| \int_{t'}^t x'(s) ds \right\| \leq M|t - t'|.$$

Hence $|t - t'| \geq \tau$. The proof of the lemma is completed. \square

Let $x \in K \setminus \text{Inv}_F(K)$; let us build a solution starting at x viable in $\overline{K \setminus \text{Inv}_F(K)}$. We know that there exists at least one solution $x(\cdot)$ that goes outside K (i.e., there exists $\tau_1 > 0$, $x(\tau_1) \in X \setminus K$). If this solution stays outside $\text{Inv}_F(K)$ then the proof is achieved. Otherwise, there exists a time $T_1 > \tau_1$ such that $x(T_1) \in (\text{Inv}_F(K) + \alpha B) \setminus \text{Inv}_F(K)$. According to Lemma 3.4, because $\|x(T_1) - x(\tau_1)\| \geq \alpha$, we have $|T_1 - \tau_1| \geq \tau$. But starting at $x(T_1)$ there exists a solution $\tilde{x}(\cdot)$ that goes outside K (there exists τ_2 , $\tilde{x}(\tau_2) \notin K$). Similar to the proof of Theorem 3.1, we obtain a solution to (1) starting at x_0 (again denoted $x(\cdot)$) such that $\tilde{x}(\tau_2) = x(T_1 + \tau_2)$. If this solution stays outside $\text{Inv}_F(K)$, then the proof is achieved, otherwise

$$\exists T_2 > T_1, x(T_2) \in (\text{Inv}_F(K) + \alpha B) \setminus \text{Inv}_F(K) \text{ with } T_2 - T_1 \geq \alpha.$$

If there is a finite number of T_i the proof is clearly achieved. If there is an infinite number of T_i this sequence converges to ∞ because $T_{n+1} - T_n \geq \alpha$. We have obtained a solution of (1) viable in $\overline{X \setminus \text{Inv}_F(K)}$. \square

When K is not compact but only closed, it is possible to prove a similar result.

PROPOSITION 3.5. *Let K be a closed set, and F a set-valued map satisfying the assumptions of Proposition 3.3.*

If $\text{Viab}_F(K) \subset \text{Int}(K)$, then $\overline{X \setminus \text{Inv}_F(K)}$ and $\partial \text{Inv}_F(K)$ are viability domains.

We then deduce a result that follows from Theorem 3.1 and Proposition 3.3.

COROLLARY 3.6. *Let F be a Lipschitzian set-valued map with nonempty convex compact values, and K a closed compact nonempty set. If $x_0 \in \partial \text{Inv}_F(K) \cap \partial \text{Viab}_F(K)$, then all solutions starting at x_0 stay on the boundary of $\text{Viab}_F(K)$, as long as they do not cross ∂K .*

Furthermore, if $\partial \text{Inv}_F(K) \cap \partial \text{Viab}_F(K) \subset \text{Int}(K)$, it is an invariance domain.

4. Backward trajectories for a differential inclusion. In previous sections, we were interested in studying solutions starting at a given point; now, we shall study solutions reaching a given point.

We compare in this section kernels associated to (1) and kernels associated to (4).

Roughly, the concatenation of solutions of (1) and (4) gives us a solution of the differential inclusion on $]-\infty, +\infty[$.

Let $\text{Viab}_{-F}(K)$ (respectively, $\text{Inv}_{-F}(K)$) denote the viability kernel of (4) (respectively, the invariance kernel) of K for the set-valued map $-F$. Of course, all results concerning boundaries of these sets are still available.

PROPOSITION 4.1. *Let F be a Lipschitzian set-valued map with nonempty convex compact values and linear growth, and K a closed set. Then*

- *The set $\text{Inv}_F(K) \cap \text{Viab}_{-F}(K)$ is an invariance domain for F .*
- *The set $\text{Inv}_{-F}(K) \cap \text{Viab}_F(K)$ is an invariance domain for $-F$.*
- *The set $\text{Viab}_F(K) \cap \text{Viab}_{-F}(K)$ is a viability domain for F and $-F$.*
- *The set $\text{Inv}_F(K) \cap \text{Inv}_{-F}(K)$ is a viability domain for F and $-F$.*

Proof. To prove this, we use a technique similar¹³ to the proof of Theorem 3.1. Let us prove, for instance, the first result.

Let x_0 belong to $\text{Inv}_F(K) \cap \text{Viab}_{-F}(K)$ and $x(\cdot)$ be a solution to (1). Fix $T > 0$. By setting $y(t) := x(T - t)$ ($t \in [0, T]$), we obtain a solution to (4) such that $y(T) = x_0 \in$

¹³ Let us recall $\text{Inv}_F(K) \subset \text{Viab}_F(K)$ and $\text{Inv}_{-F}(K) \subset \text{Viab}_{-F}(K)$.

$\text{Viab}_{-F}(K)$. Hence, there exists $\tilde{y}(\cdot)$ a solution to (4) starting at x_0 and viable in K with respect to $-F$. The concatenation of $y([0, T])$ and $\tilde{y}([T, \infty[)$ provides a solution starting at $x(T)$ viable in K (with respect to $-F$). Hence $x(T) \in \text{Viab}_{-F}(K)$; since T is arbitrary, the proof follows. \square

PROPOSITION 4.2. *If assumptions of Proposition 4.1 hold true, then*

- (i) $\text{Viab}_F(K) \subset \text{Int}(K) \Rightarrow \text{Viab}_F(K) \subset \text{Inv}_{-F}(K)$
- (ii) $\text{Viab}_{-F}(K) \subset \text{Int}(K) \Rightarrow \text{Viab}_{-F}(K) \subset \text{Inv}_F(K)$.

It follows by exactly the same arguments as in the previous proof.

Remark. We can now “mix” all these subsets and kernels to prove results of the type

$$\begin{aligned}\text{Viab}_F(K) \cap \text{Viab}_F(\hat{K}) &\subset K^b, \\ \text{Viab}_F(K) \cap \text{Inv}_{-F}(\hat{K}) &\subset K^b \cap \hat{K}^{i-},\end{aligned}$$

and so on. \square

5. An application to control systems with one target: Semipermeable barriers. A question naturally arises: Why is it useful to study the boundary of viability or invariance kernels? We give an example of controlled system with one target.

5.1. Certain and possible victory and defeat domains. We can model the controlled system¹⁴

$$(5) \quad x'(t) = f(x(t), u(t)), \quad u(t) \in U(x(t))$$

through the differential inclusion (1) by setting $F(x) := f(x, U(x))$. We shall assume that F is Lipschitzian with convex compact nonempty values.¹⁵

Our problem is to drive in finite time the state x inside a given open set C starting at a point outside of C . This has a precise mathematical sense by using the viability and invariance kernels of $K := X \setminus \overline{C}$. Let us introduce some definition of victory and defeat domains (see [3]).

DEFINITION 5.1. We define

- the domain of *certain defeat* by the set $\text{Inv}_F(X \setminus C)$,
- the domain of *possible defeat* by the set $\text{Viab}_F(X \setminus C)$,
- the domain of *certain victory* by the set $\overline{K \setminus \text{Viab}_F(X \setminus C)}$,
- the domain of *possible victory* by the set $\overline{K \setminus \text{Inv}_F(X \setminus C)}$.

Let us make more precise the qualitative behaviour of solutions in these domains.

PROPOSITION 5.2. *If $x_0 \in \text{Inv}_F(X \setminus C)$, then no solution to (5), starting at x_0 , can reach C (certain defeat).*

If $x_0 \in \text{Viab}_F(X \setminus C)$, there exist solutions of (5), starting at x_0 , that never reach C (possible defeat).

If $x_0 \in X \setminus \text{Viab}_F(X \setminus C)$, then all solutions of (5), starting at x_0 , reach C in finite time (certain victory).

If $x_0 \in K \setminus \text{Inv}_F(X \setminus C)$, there exist solutions to (5), starting at x_0 , that reach C in finite time (possible victory).

Proof. It is the obvious consequence of Definitions 1.2 and 1.1. \square

5.2. Semipermeable barrier. We shall define some subsets of the boundaries of these victory and defeat domains.

DEFINITION 5.3. The *barrier* is the set

$$\partial \text{Inv}_F(X \setminus C) \setminus \partial C.$$

¹⁴ Results of this paper can be easily extended to the nonautonomous case, i.e., $x'(t) = f(t, x(t), u(t))$, $u(t) \in U(t, x(t))$ (see [25]).

¹⁵ In particular, it is satisfied if f is Lipschitzian affine with respect to the control and U is a Lipschitzian set-valued map with nonempty convex compact values.

The *strict barrier* is the set

$$\partial \text{Viab}_F(X \setminus C) \setminus \partial C.$$

We can note that the barrier is contained in the intersection of the certain defeat domain and the possible victory domain. The strict barrier is contained in the intersection of the possible defeat domain and the certain victory domain. We can translate the results of § 3, and so we have a qualitative description of the behaviour of solution on the barriers.

PROPOSITION 5.4. *The strict barrier is a local viability domain.¹⁶ Furthermore, all solutions starting at any state x_0 of the strict barrier that are viable in $X \setminus C$ remain in this set until it reaches \bar{C} (and there exists such solution).*

The barrier is a viability domain¹⁷ as soon as

$$\bar{C} \cap \text{Viab}_F(X \setminus C) = \emptyset.$$

Proof. The first result is a consequence of Theorem 3.1, and the second one is a consequence of Proposition 3.5. \square

This generalizes the concept of semipermeable barriers (see [9], [10]). Recall that a C^1 -surface is semipermeable when it satisfies an equation such that $\max_u f(x, u) \cdot n \leq 0$ or $\min_u f(x, u) \cdot n \geq 0$ (where n is the normal vector of the surface). It means that the solutions of (5) are able to cross the surface in *only one direction*. Let us make this idea more precise by using the partition of § 2.

DEFINITION 5.5. Let A be a closed set. A subset B of ∂A is semipermeable for A if and only if, for any point x_0 of B , any solution $x(\cdot)$ starting at x_0 is locally viable in A .

Remark. Let us note that an obvious consequence of this definition is $B \cap A^e = \emptyset$.

Thanks to Proposition 5.4, we can state the following proposition.

PROPOSITION 5.6. *The strict barrier is semipermeable for $\text{Viab}_F(X \setminus C)$.*

In fact, thanks to Proposition 5.4, we know that a solution of (5) cannot cross the strict barrier if it comes from the exterior of the viability kernel $\text{Viab}_F(X \setminus C)$, but the converse is possible (i.e., there could exist solutions coming from the interior of the kernel that cross the strict barrier).

We can note that, thanks to Corollary 3.6, the intersection of the strict barrier and the barrier is a *local invariance domain*¹⁸ (if it is nonempty).

Acknowledgment. I am indebted to Halina Frankowska for her help and advice.

REFERENCES

- [1] J. P. AUBIN AND A. CELLINA, *Differential inclusions*, Grundlehren Math. Wiss., 264 (1984).
- [2] J. P. AUBIN AND I. EKELAND, *Applied Nonlinear Analysis*, Wiley-Interscience, New York, 1984.
- [3] J. P. AUBIN, *Differential games: A viability approach*, SIAM J. Control Optim., 28 (1990), pp. 1–27.
- [4] ———, *A survey of viability theory*, SIAM J. Control Optim., 28 (1990), pp. 749–788.
- [5] ———, *Viability Theory*, to appear.
- [6] J. P. AUBIN AND H. FRANKOWSKA, *Set-Valued Analysis*, Birkhäuser, Boston, 1990.
- [7] ———, *Viability kernels of control systems*, in Proc. Conference Nonlinear Synthesis Progress in Systems and Controls, Birkhäuser, Boston, to appear.

¹⁶ A set K is a local viability domain if and only if, starting at any point of K , there exists at least one solution locally viable in K .

¹⁷ When we assume that $X \setminus C$ is compact, according to Proposition 3.3, we have the same conclusion if $\text{Inv}_F(X \setminus C) \cap \bar{C} = \emptyset$.

¹⁸ A set K is a local invariance domain if and only if all solutions starting at any point of K are locally viable in K .

- [8] C. BERGE, *Espaces topologiques, Fonctions Multivoques*, Dunod, Paris, 1961.
- [9] P. BERNHARD, *Commande Optimale, Décentralisation, et Jeux Dynamiques*, Dunod, Paris, 1976.
- [10] ———, *Differential games*, in *Systems and Control Encyclopedia*, Theory Technology Application, M. G. Singh, ed., Pergamon Press, Oxford, UK, 1988, pp. 1004–1023.
- [11] P. BERNHARD AND B. LARROUTUROU, *Etude de la barrière pour un problème de fuite optimale dans le plan*, preprint, Rapport de recherche INRIA, Le Chesnay, France, 1989.
- [12] O. DORDAN, *Analyse qualitative*, thesis, Paris, France, University Paris IX, 1990.
- [13] A. F. FILIPPOV, *On some problems of optimal control theory*, *Vestnik Moskov. Univ. Math.*, 2 (1958), pp. 25–32. (English translation in *SIAM J. Control*, 1 (1962), pp. 76–84.)
- [14] H. FRANKOWSKA AND M. QUINCAMPOIX, *Viability kernels of differential inclusions with constraints: Algorithm and applications*, *Math. Systems Estimation Control*, 1 (1991) pp. 371–388.
- [15] ———, *Un algorithme déterminant les noyaux de viabilité pour les inclusions différentielles avec contraintes*, *Comptes Rendus de l'Académie des Sciences, Série I, PARIS*, t. 312 (1991), pp. 31–36.
- [16] H. GUSEINOV, A. I. SUBBOTIN, AND V. N. USHAKOV, *Derivatives for multivalued mappings with applications to game theoretical problems of control*, *Problems Control Inform. Theory*, 14 (1985), pp. 155–167.
- [17] G. HADDAD, *Monotone trajectories of differential inclusions with memory*, *Israel J. Math.*, 39 (1981), pp. 38–100.
- [18] R. ISAACS, *Differential Games*, John Wiley, New York, 1965.
- [19] V. KRIVAN, *Construction of population growth equations in the presence of viability constraints*, *J. Math. Biol.*, to appear.
- [20] N. N. KRASOVSKI, *The Control of a Dynamic System*, Nauka, Moscow, 1986.
- [21] A. B. KURZHANSKII, *Control and Observation under Conditions of Uncertainty*, Nauka, Moscow, 1977.
- [22] ———, *On the analytical properties of viability tubes of trajectories of differential systems*, *Dokl. Acad. Nauk SSSR*, 287 (1986), pp. 1047–1050.
- [23] A. B. KURZHANSKII AND T. F. FILIPPOVA, *On a description of the set of viable trajectories of a differential inclusion*, *Soviet. Math. Dokl.*, 34 (1987).
- [24] M. QUINCAMPOIX, *Frontières de domaines d'invariance et de viabilité pour des inclusions différentielles avec contraintes*, *Comptes Rendus de l'Académie des Sciences, Série I, PARIS*, t. 311 (1990), pp. 904–914.
- [25] ———, *Playable differential games*, *J. Math. Anal. Appl.*, to appear.
- [26] E. O. ROXIN, *Stability in general control systems*, *J. Differential Equations*, 1 (1965), pp. 115–150.
- [27] ———, *On stability in control systems*, *SIAM J. Control*, 3 (1965), pp. 357–372.
- [28] ———, *On general dynamical systems defined by contingent equations*, *J. Differential Equations*, 1 (1965), pp. 188–205.

DYNAMIC DISTURBANCE DECOUPLING FOR NONLINEAR SYSTEMS*

H. J. C. HUIJBERTS[†], H. NIJMEIJER[†], AND L. L. M. VAN DER WEGEN[‡]

Abstract. In analogy with the dynamic input-output decoupling problem the dynamic disturbance decoupling problem for nonlinear systems is introduced. A local solution of this problem is obtained in the case that the system under consideration is invertible. The solution is given in algebraic as well as in geometric terms. The theory is illustrated by means of two examples: a mathematical one and an example of a voltage frequency controlled induction motor.

Key words. nonlinear control systems, dynamic disturbance decoupling, invertibility, dynamic precompensation

AMS(MOS) subject classifications. 93C10, 93B50, 93C35

1. Introduction. Consider a nonlinear multi-input-multi-output control system Σ of the form

$$(1) \quad \Sigma \begin{cases} \dot{x} = f(x) + g(x)u + p(x)q \\ y = h(x), \end{cases}$$

where $x \in \mathcal{X}$, an open subset of \mathbb{R}^n , the inputs $u \in \mathbb{R}^m$, the outputs $y \in \mathbb{R}^p$, the disturbances $q \in \mathbb{R}^r$, f and h are vector-valued analytic functions, and g and p are matrix-valued analytic functions, all of appropriate dimensions. In the disturbance decoupling problem (DDP) for (1), we search for a regular static state feedback

$$(2) \quad u = \alpha(x) + \beta(x)v,$$

with v a new m -dimensional control and $\beta(x)$ a nonsingular $m \times m$ matrix for all x , so that in the feedback modified dynamics

$$(3) \quad \dot{x} = f(x) + g(x)\alpha(x) + g(x)\beta(x)v + p(x)q,$$

the disturbances q do not affect the outputs y . A local solution of the DDP using differential geometric tools was initiated in [13] and [9] and has led to a more or less complete understanding of this problem; see, e.g., [12], [18]. The nonlinear DDP forms a direct generalization of the linear DDP, and the theory about the nonlinear DDP typically extends the well-known linear geometric theory (cf. [23]) to a nonlinear context.

The purpose of this paper is to study a dynamic version of the DDP for the nonlinear system (1). That is, instead of a static feedback law (2) we allow for a *regular* dynamic state feedback

$$(4) \quad \begin{aligned} \dot{z} &= \alpha(x, z) + \beta(x, z)v, \\ u &= \gamma(x, z) + \delta(x, z)v, \end{aligned}$$

with z the μ -dimensional compensator state and v an m -dimensional new control, and the regularity of (4) means that the system (4) with inputs v and outputs u is invertible for all z and constant x . Note that a somewhat different definition of regular

* Received by the editors January 10, 1990; accepted for publication (in revised form) January 2, 1991.

[†] Department of Applied Mathematics, University of Twente, Post Office Box 217, 7500 AE Enschede, the Netherlands.

[‡] School of Management Studies, University of Twente, Post Office Box 217, 7500 AE Enschede, the Netherlands.

dynamic state feedback was given in [15]. In the dynamic disturbance decoupling problem (DDDP), we require that in the modified dynamics

$$(5) \quad \begin{aligned} \dot{x} &= f(x) + g(x)\gamma(x, z) + g(x)\delta(x, z)v + p(x)q, \\ \dot{z} &= \alpha(x, z) + \beta(x, z)v \end{aligned}$$

the disturbances q do not influence the outputs y . Clearly, the static DDP forms a special case of the DDDP by assuming that $\mu = 0$. As noted before, the theory for solving the nonlinear DDP is very much based on a proper extension of the solution of the linear DDP. We are therefore led to think that similarly a solution of the nonlinear DDDP naturally extends the DDDP for linear systems. However, a very simple argument shows that for linear systems the DDDP is solvable if and only if the static DDP is solvable (see, e.g., [1], [2]). Although an analogous result is also true for single-output nonlinear systems, i.e., when $p = 1$, this conclusion no longer holds true for multi-output nonlinear systems. In other words, when $p > 1$ it may happen that the nonlinear DDDP is (locally) solvable, whereas the nonlinear DDP is not.

Our goal is to establish necessary and sufficient conditions for the solvability of the DDDP, thereby discussing various different algebraic and geometric aspects of this problem for the case that the system (1) with $q \equiv 0$ is square and invertible.

The organization of the paper is as follows. In § 2 we introduce the dynamic disturbance decoupling problem with disturbance measurements (DDDPdm) and the dynamic disturbance decoupling problem (DDDP), and we show that both problems are locally solvable if and only if they are solvable by means of a compensator that is obtained from Singh's algorithm. In §§ 3 and 4 we translate the conditions for solvability of the DDDPdm obtained in § 2 into intrinsic and algorithm-independent conditions, using differential algebraic and geometric tools, respectively. In § 5 the theory of the foregoing sections will be applied to an example of a voltage frequency controlled induction motor as was described in [3]. Section 6 contains the conclusions of the paper.

2. The dynamic disturbance decoupling problem (DDDP). In this section we formulate and solve two kinds of DDDPs. These problems are dynamic extensions of the well-known static state feedback DDP, respectively, the static state feedback DDP with disturbance measurements.

DEFINITION 2.1. Consider the analytic system Σ and let a point $x_0 \in \mathcal{X}$ be given.

1. The DDDP is said to be locally solvable around x_0 if there exist an analytic dynamic state feedback for Σ of the form (4), to be denoted as R , with $z \in \mathbb{R}^\mu$, a neighborhood $U \subset \mathcal{X}$ of x_0 , and an open subset $\mathcal{Z} \subset \mathbb{R}^\mu$ such that R , with inputs v and outputs u , is invertible for all constants $x \in U$ and $z \in \mathcal{Z}$ and the outputs of the composite system $\Sigma \circ R$ restricted to $U \times \mathcal{Z}$ are independent of the disturbances.

2. The dynamic disturbance decoupling problem with disturbance measurements (DDDPdm) is said to be locally solvable around x_0 if there exist a dynamic state feedback for Σ of the form

$$(6) \quad Q \begin{cases} \dot{z} = \alpha(x, q, z) + \beta(x, q, z)v \\ u = \gamma(x, q, z) + \delta(x, q, z)v, \end{cases}$$

with $z \in \mathbb{R}^\mu$, a neighborhood $U \subset \mathcal{X}$ of x_0 , and an open subset $\mathcal{Z} \subset \mathbb{R}^\mu$, such that (6) with inputs v and outputs u is invertible for all constant $x \in U$ and all $q \in \mathbb{R}^r$ and $z \in \mathcal{Z}$, and the outputs of the composite system $\Sigma \circ Q$ restricted to $U \times \mathcal{Z}$ are independent of the disturbances.

If we furthermore require the compensators Q and R to be static state feedback compensators (i.e., $\mu = 0$), the problem will be referred to by DDPdm and DDP, respectively.

Recall that the DDP is locally solvable if and only if $\mathcal{P} \subset \Delta^*$ and that the DDPdm is locally solvable if and only if $\mathcal{P} \subset \Delta^* + \mathcal{G}$, where $\mathcal{G} = \text{span}\{g_1, \dots, g_m\}$, $\mathcal{P} = \text{span}\{p_1, \dots, p_r\}$, and Δ^* is the maximal locally controlled invariant distribution contained in $\ker dh$ (cf. [13]). It is well known (see, e.g., [1], [2]) that for linear systems the DDDP is solvable if and only if the DDP is solvable. That this is not the case for nonlinear systems can be seen from the following example.

Example 2.2. Consider the nonlinear system

$$\begin{aligned}
 \dot{x}_1 &= x_2 u_1, & y_1 &= x_1, \\
 \dot{x}_2 &= x_5, & y_2 &= x_3, \\
 \dot{x}_3 &= x_2 + x_4 + x_4 u_1, \\
 \dot{x}_4 &= u_2, \\
 \dot{x}_5 &= x_1 u_1 + q.
 \end{aligned}
 \tag{7}$$

For this system we have $\Delta^* = \{0\}$. Hence the DDP is not locally solvable. However, if we apply the compensator

$$\begin{aligned}
 \dot{z} &= v_1, \\
 u_1 &= z, \\
 u_2 &= v_2,
 \end{aligned}
 \tag{8}$$

where v_1, v_2 are the new inputs, we find for the compensated system (7), (8) that

$$\Delta_e^* = \text{span} \left\{ \frac{\partial}{\partial x_5}, x_2(1+z) \frac{\partial}{\partial x_2} + (x_2 - zx_4) \frac{\partial}{\partial x_4} - z(1+z) \frac{\partial}{\partial z} \right\},$$

where Δ_e^* is the maximal locally controlled invariant distribution contained in $\ker dh$ for (7), (8). Hence it is clear that the DDP for (7), (8) is solvable and thus the DDDP is solvable for (7).

In the following, we make the standing assumption, below.

(A1). The system Σ is square, i.e., $p = m$.

Instrumental in the solution of the DDDPdm is what we like to call a Singh compensator, which can be obtained via the so-called Singh algorithm. Singh's algorithm has been introduced in [19] for calculation of a left inverse of a nonlinear system. It is a generalization of the algorithm from [8], which was only applicable under some restrictive assumptions. Let Σ_0 denote the system Σ with $q \equiv 0$. We briefly repeat Singh's algorithm for the system Σ_0 , following [5].

ALGORITHM 2.3. Consider the analytic nonlinear system Σ_0 , satisfying (A1). Let a point $x_0 \in \mathcal{X}$ be given.

Step 1.

Calculate

$$\dot{y} = \frac{\partial h}{\partial x} [f(x) + g(x)u] =: a_1(x) + b_1(x)u
 \tag{9}$$

and assume that $b_1(x)$ has full rank ρ_1 on a neighborhood of x_0 . Define $s_1 := \rho_1$. Permute, if necessary, the components of the output, so that the first ρ_1 rows of $b_1(x)$ are linearly independent. Decompose y according to

$$(10) \quad \dot{y} = \begin{pmatrix} \dot{\hat{y}}_1 \\ \dot{\hat{y}}_1 \end{pmatrix},$$

where $\dot{\hat{y}}_1$ consists of the first ρ_1 rows of \dot{y} . Since the last rows of $b_1(x)$ are linearly dependent on the first ρ_1 rows, we can write

$$(11) \quad \dot{\hat{y}}_1 = \tilde{a}_1(x) + \tilde{b}_1(x)u, \quad \dot{\hat{y}}_1 = \hat{y}_1(x, \dot{\hat{y}}_1),$$

where the last equation is affine in $\dot{\hat{y}}_1$. Finally, set $\tilde{B}_1(x) := \tilde{b}_1(x)$.

Step $k+1$.

Suppose that in Steps 1 through k , $\dot{\hat{y}}_1, \dots, \tilde{y}_k^{(k)}, \hat{y}_k^{(k)}$ have been defined so that

$$(12) \quad \begin{aligned} \dot{\hat{y}}_1 &= \tilde{a}_1(x) + \tilde{b}_1(x)u, \\ &\vdots \\ \tilde{y}_k^{(k)} &= \tilde{a}_k(x, \{\tilde{y}_i^{(j)} \mid 1 \leq i \leq k-1, i \leq j \leq k\}) \\ &\quad + \tilde{b}_k(x, \{\tilde{y}_i^{(j)} \mid 1 \leq i \leq k-1, i \leq j \leq k-1\})u, \\ \hat{y}_k^{(k)} &= \hat{y}_k^{(k)}(x, \{\tilde{y}_i^{(j)} \mid 1 \leq i \leq k, i \leq j \leq k\}). \end{aligned}$$

Suppose also that there exist $\tilde{y}_{i0}^{(j)}$ ($1 \leq i \leq k-1, i \leq j \leq k-1$) such that the matrix $\tilde{B}_k := [\tilde{b}_1^T, \dots, \tilde{b}_k^T]^T$ has full rank ρ_k on a neighborhood of $(x_0, \{\tilde{y}_{i0}^{(j)} \mid 1 \leq i \leq k-1, i \leq j \leq k-1\})$. Then calculate

$$(13) \quad \hat{y}_k^{(k+1)} = \frac{\partial}{\partial x} \hat{y}_k^{(k)}[f(x) + g(x)u] + \sum_{i=1}^k \sum_{j=i}^k \frac{\partial \hat{y}_k^{(k)}}{\partial \tilde{y}_i^{(j)}} \tilde{y}_i^{(j+1)}$$

and write it as

$$(14) \quad \begin{aligned} \hat{y}_k^{(k+1)} &= a_{k+1}(x, \{\tilde{y}_i^{(j)} \mid 1 \leq i \leq k, i \leq j \leq k+1\}) \\ &\quad + b_{k+1}(x, \{\tilde{y}_i^{(j)} \mid 1 \leq i \leq k, i \leq j \leq k\})u. \end{aligned}$$

Define $B_{k+1} := [\tilde{B}_k^T, b_{k+1}^T]^T$, and suppose that there exist $\tilde{y}_{i0}^{(j)}$ ($1 \leq i \leq k, i \leq j \leq k$) such that B_{k+1} has constant rank ρ_{k+1} on a neighborhood of $(x_0, \{\tilde{y}_{i0}^{(j)} \mid 1 \leq i \leq k, i \leq j \leq k\})$. Permute, if necessary, the components of $\hat{y}_k^{(k+1)}$ so that on this neighborhood the first ρ_{k+1} rows of B_{k+1} are linearly independent. Decompose $\hat{y}_k^{(k+1)}$ as

$$\hat{y}_k^{(k+1)} = (\hat{y}_{k+1}^{(k+1)T} \quad \hat{y}_{k+1}^{(k+1)T})^T,$$

where $\hat{y}_{k+1}^{(k+1)}$ consists of the first $s_{k+1} := (\rho_{k+1} - \rho_k)$ rows. Since the last rows of B_{k+1} are linearly dependent on the first ρ_{k+1} rows, we can write

$$(15) \quad \begin{aligned} \dot{\hat{y}}_1 &= \tilde{a}_1(x) + \tilde{b}_1(x)u, \\ &\vdots \\ \tilde{y}_{k+1}^{(k+1)} &= \tilde{a}_{k+1}(x, \{\tilde{y}_i^{(j)} \mid 1 \leq i \leq k, i \leq j \leq k+1\}) \\ &\quad + \tilde{b}_{k+1}(x, \{\tilde{y}_i^{(j)} \mid 1 \leq i \leq k, i \leq j \leq k\})u, \\ \hat{y}_{k+1}^{(k+1)} &= \hat{y}_{k+1}^{(k+1)}(x, \{\tilde{y}_i^{(j)} \mid 1 \leq i \leq k+1, i \leq j \leq k+1\}). \end{aligned}$$

Finally, set $\tilde{B}_{k+1} := [\tilde{B}_k^T, \tilde{b}_{k+1}^T]^T$.

It should be noted that the integers $\rho_1, \dots, \rho_k, \dots$, defined above, do not depend on the particular permutation of the rows of $\hat{y}_k^{(k+1)}$ we employ, cf. [5]. So, using the

algorithm we obtain a uniquely defined sequence of integers $0 \leq \rho_1 \leq \dots \leq \rho_k \leq \dots \leq m$. The integer $\rho^* := \rho_n$ is called the *rank* of the system (1), cf. [18], [5]. We associate a notion of regularity with Singh's algorithm in the following way. (See [4] for a slightly different notion of regularity.)

DEFINITION 2.4. Let a point $x_0 \in \chi$ be given. We call x_0 a *strongly regular point* for Σ if, for each application of Singh's algorithm to Σ_0 , the constant rank assumptions of the algorithm are satisfied.

Besides (A1), we also introduce the following assumption.

(A2). The system Σ_0 is invertible, i.e., $\rho^* = m$.

Consider a system Σ satisfying (A1), (A2). Then we define a Singh compensator for Σ as follows (see also [20]). Let x_0 be a strongly regular point for Σ and apply Singh's algorithm for Σ_0 . This yields at the n th step

$$(16) \quad \begin{aligned} \dot{\tilde{y}}_n &= \tilde{A}_n(x, \{\tilde{y}_i^{(j)} | 1 \leq i \leq n-1, i \leq j \leq n\}) \\ &\quad + \tilde{B}_n(x, \{\tilde{y}_i^{(j)} | 1 \leq i \leq n-1, i \leq j \leq n-1\})u, \end{aligned}$$

where $\tilde{Y}_n = (\tilde{y}_1^T, \dots, \tilde{y}_n^{(n-1)T})$ and where \tilde{B}_n is invertible on a neighborhood of $(x_0, \{\tilde{y}_{i0}^{(j)} | 1 \leq i \leq n-1, i \leq j \leq n-1\})$ for some $\tilde{y}_{i0}^{(j)} (1 \leq i \leq n, i \leq j \leq n)$. Then (16) yields on this neighborhood

$$(17) \quad u = \tilde{B}_n^{-1}[\dot{\tilde{Y}}_n - \tilde{A}_n].$$

For $i = 1, \dots, m$, let γ_i be the lowest time-derivative and δ_i be the highest time-derivative of y_i appearing in (17). Then we rewrite (17) as

$$(18) \quad \begin{aligned} u &= \phi_1(x, \{y_i^{(j)} | 1 \leq i \leq m, \gamma_i \leq j \leq \delta_i - 1\}) \\ &\quad + \sum_{i=1}^m \phi_{2i}(x, \{y_i^{(j)} | 1 \leq i \leq m, \gamma_i \leq j \leq \delta_i - 1\})y_i^{(\delta_i)} \end{aligned}$$

for certain locally analytic vector-valued functions $\phi_1, \phi_{2i} (i = 1, \dots, m)$.

Let $z_i (i = 1, \dots, m)$ be a vector of dimension $\delta_i - \gamma_i$ and consider the system

$$(19) \quad \begin{aligned} \dot{z}_i &= A_i z_i + B_i v_i (i = 1, \dots, m), \\ u &= \phi_1(x, z_1, \dots, z_m) + \sum_{i=1}^m \phi_{2i}(x, z_1, \dots, z_m)v_i, \end{aligned}$$

with inputs v_1, \dots, v_m , outputs u , (A_i, B_i) in Brunovsky canonical form, and $z_{i0} = (y_{i0}^{(\gamma_i)}, \dots, y_{i0}^{(\delta_i-1)})^T (i = 1, \dots, m)$. Then (19) is called a *Singh compensator* for Σ around x_0 . Note that the Singh compensator (19) is an inverse of the system Σ_0 around x_0 .

Remark 2.5. The Singh compensator, constructed above, has dimension $\sigma = \sum_{i=1}^m (\delta_i - \gamma_i)$. It can be shown (see [11]) that every Singh compensator has dimension σ . Moreover, the numbers $\delta_i (i = 1, \dots, m)$ are equal to the essential orders (cf. [7]) of Σ (see also [11]).

We obtain a Singh compensator with disturbance feedthrough for Σ around x_0 by extending (19) in the following way. Define in the above notation

$$(20) \quad \phi_3(x, \{y_i^{(j)} | 1 \leq i \leq m, \gamma_i \leq j \leq \delta_i - 1\}) := -\tilde{B}_n^{-1} \frac{\partial \tilde{Y}_n}{\partial x} p(x)$$

and consider the following extension of (19):

$$(21) \quad \begin{aligned} \dot{z}_i &= A_i z_i + B_i v_i (i = 1, \dots, m), \\ u &= \phi_1(x, z_1, \dots, z_m) + \sum_{i=1}^m \phi_{2i}(x, z_1, \dots, z_m)v_i + \phi_3(x, z_1, \dots, z_m)q. \end{aligned}$$

Then (21) is called a *Singh compensator with disturbance feedthrough* for Σ around x_0 . It can be shown that the Singh compensator as well as the Singh compensator with disturbance feedthrough constitute a regular dynamic state feedback (cf. [10]).

We will now state our main result. In this statement we employ the following notation. If we apply Singh's algorithm to Σ_0 , the $\hat{y}_k^{(k)} (k=0, \dots, n; \hat{y}_0 := y)$ can be viewed as functions on $\mathcal{X}_e := \mathcal{X} \times \mathbb{R}^{nm}$. By the same token, $\ker d\hat{y}_k^{(k)} (k=0, \dots, n)$ defines a distribution on \mathcal{X}_e . Define the distributions $\mathcal{G}_e, \mathcal{P}_e$ on \mathcal{X}_e by $\mathcal{G}_e := \mathcal{G} \times \{0\}, \mathcal{P}_e := \mathcal{P} \times \{0\}$.

THEOREM 2.6. *Consider the analytic system Σ . Assume that it satisfies (A1), (A2) and that x_0 is a strongly regular point for Σ .*

1. (a) *The DDDP is locally solvable around x_0 if and only if it is solvable via a Singh compensator for Σ around x_0 .*

(b) *The above condition is equivalent to the condition that for each application of Singh's algorithm to Σ_0 we have for $k=0, \dots, n-1$:*

$$(22) \quad \mathcal{P}_e \subset \ker d\hat{y}_k^{(k)}.$$

2. (a) *The DDDPdm is locally solvable around x_0 if and only if it is solvable via a Singh compensator with disturbance feedthrough for Σ around x_0 .*

(b) *the above condition is equivalent to the condition that for each application of Singh's algorithm to Σ_0 we have, for $k=0, \dots, n-1$,*

$$(23) \quad \mathcal{P}_e \subset \ker d\hat{y}_k^{(k)} + \mathcal{G}_e.$$

Proof. We will prove only part 2. The proof of part 1 is analogous.

Sufficiency. Consider Σ and assume that it satisfies (A1), (A2), and let x_0 be a strongly regular point for Σ . Assume that for each application of Singh's algorithm to Σ_0 (23) holds for $k=0, \dots, n-1$. Apply Singh's algorithm to Σ_0 around x_0 , yielding a reordering $\tilde{y}_1, \dots, \tilde{y}_n$ of the outputs. Note that, without loss of generality, we may assume that for $k=1, \dots, n$: $\hat{y}_k = (\tilde{y}_{k+1}^T, \dots, \tilde{y}_n^T)^T$. The first step of Singh's algorithm applied to Σ_0 around x_0 yields

$$(24) \quad \begin{aligned} \dot{\tilde{y}}_1 &= \tilde{a}_1(x) + \tilde{b}_1(x)u, \\ \dot{\hat{y}}_1 &= \hat{a}_1(x) + \hat{b}_1(x)u, \end{aligned}$$

where $\tilde{b}_1(x)$ has full row rank ρ_1 on a neighborhood U of x_0 in \mathcal{X} . Let $\tilde{b}_1^+(x)$ be a right inverse of $\tilde{b}_1(x)$ on U . Then on U , \hat{y}_1 takes the form

$$(25) \quad \hat{y}_1 = \hat{a}_1(x) + \hat{b}_1(x)\hat{b}_1^+(x)(\dot{\tilde{y}}_1 - \tilde{a}_1(x)).$$

Note that, since on U each of the rows of $\hat{b}_1(x)$ is linearly dependent on the rows of $\tilde{b}_1(x)$, the form of \hat{y}_1 is independent of the choice of $\tilde{b}_1^+(x)$. For Σ we have

$$(26) \quad \begin{aligned} \dot{\tilde{y}}_1 &= \frac{\partial \tilde{y}_1}{\partial x} [f(x) + g(x)u + p(x)q] =: \tilde{a}_1(x) + \tilde{b}_1(x)u + \tilde{c}_1(x)q, \\ \dot{\hat{y}}_1 &= \frac{\partial \hat{y}_1}{\partial x} [f(x) + g(x)u + p(x)q] =: \hat{a}_1(x) + \hat{b}_1(x)u + \hat{c}_1(x)q, \end{aligned}$$

with $\tilde{a}_1(x), \tilde{b}_1(x), \hat{a}_1(x), \hat{b}_1(x)$ as in (24). It can easily be checked that the fact that (23) holds for $k=0$ is equivalent to the existence of a $\sigma_1(x)$ such that

$$(27) \quad \tilde{b}_1(x)\sigma_1(x) = \tilde{c}_1(x), \hat{b}_1(x)\sigma_1(x) = \hat{c}_1(x).$$

Then (26) and (27) yield

$$(28) \quad \dot{\hat{y}}_1 = \hat{a}_1(x) + \hat{b}_1(x)\tilde{b}_1^+(x)(\dot{\tilde{y}}_1 - \tilde{a}_1(x)).$$

Hence for Σ , \hat{y}_1 is given by the same expression as for Σ_0 . Applying the above arguments repeatedly, we can show that for Σ $\hat{y}_k^{(k)}$ ($k=1, \dots, n$) has the same form as for Σ_0 , and that for Σ , \hat{Y}_n takes the form (see also (16))

$$(29) \quad \dot{\hat{Y}}_n = \tilde{A}_n \hat{Y}_n + \tilde{B}_n u + \frac{\partial \hat{Y}_n}{\partial x} p(x) q.$$

This implies that if we apply the Singh compensator with disturbance feedthrough to Σ , the outputs of the resulting system satisfy

$$(30) \quad \begin{aligned} \frac{\partial y_i^{(j)}}{\partial q} &= 0 \quad (1 \leq i \leq m, 0 \leq j \leq \delta_i - 1), \\ y_i^{(\delta_i)} &= v_i. \end{aligned}$$

Hence the Singh compensator with disturbance feedthrough locally solves the DDDPdm for Σ around x_0 .

Necessity. Assume that the DDDPdm is locally solvable around x_0 by means of a compensator Q of the form (6). Apply Singh's algorithm to Σ_0 around x_0 , yielding a reordering $\tilde{y}_1, \dots, \tilde{y}_n$ of the outputs. Then with the notation of the necessity part of this proof, we have in particular for Σ

$$(31) \quad \begin{aligned} \dot{\hat{y}}_1 &= \tilde{a}_1(x) + \tilde{b}_1(x)u + \tilde{c}_1(x)q, \\ \dot{\hat{y}}_1 &= \hat{a}_1(x) + \hat{b}_1(x)\tilde{b}_1^+(x)(\dot{\hat{y}}_1 - \tilde{a}_1(x)) + (\hat{c}_1(x) - \hat{b}_1(x)\tilde{b}_1^+(x)\tilde{c}_1(x))q. \end{aligned}$$

Assume that (23) does not hold for $k=0$. This implies that $d_1(x) := \hat{c}_1(x) - \hat{b}_1(x)\tilde{b}_1^+(x)\tilde{c}_1(x) \neq 0$. Since Q solves the DDDPdm for Σ , the q -dependence in (31) should disappear if we plug the output of Q in (31). From the form of (31), it is clear that this is only possible if Q imposes the constraint $d_1(x)=0$ on the system $\Sigma \circ Q$. However, this would imply that the DDDPdm is at most solvable on a neighborhood of x_0 in $\{x \mid d_1(x)=0\}$ and not on a neighborhood of x_0 in \mathcal{X} . Hence we must necessarily have that (23) holds for $k=0$. Next assume that (23) does not hold for $k=1$. Then by the same arguments as above we will have for Σ that $\hat{y}_2^{(2)}$ explicitly depends on q via a function $d_2(x, \hat{y}_1)$ and that this q -dependence can only disappear if Q imposes the constraint $d_2(x, \hat{y}_1)=0$ on the system. However, by Lemma 1 of [14], this would imply that the rank of $\Sigma \circ Q$ is smaller than the rank of Σ , which would contradict the invertibility of Q . Therefore (23) must hold for $k=1$. Applying this argument repeatedly, we show that (23) holds for $k=0, \dots, n-1$, which establishes our claim. By the sufficiency part of this proof, this also immediately implies that we can solve the DDDPdm around x_0 via a Singh compensator with disturbance feedthrough. \square

Example 2.2 (continued). The Singh compensator

$$(32) \quad \begin{aligned} \dot{z} &= v_1, \\ u_1 &= \frac{z}{x_2}, \\ u_2 &= \frac{x_5(x_4 z - x_2)}{x_2 + z} - \frac{x_4}{x_2 + z} v_1 + \frac{x_2}{x_2 + z} v_2, \end{aligned}$$

solves the DDDP for system (7) of the earlier part of Example 2.2.

Remark 2.7. It is easily seen that if the input-output decoupling problem for Σ is locally solvable by means of a static state feedback, then the DDDPdm is locally

solvable if and only if the DDDPdm is. This implies in particular that if the output of Σ is one-dimensional, then the DDDPdm is locally solvable if and only if the DDPdm is.

3. Algebraic conditions. In this section we translate the conditions for solvability of the DDDPdm and the DDDP obtained in § 2 into intrinsic and algorithm-independent conditions. It will turn out that the conditions can be stated in terms of a certain structure at infinity.

For nonlinear systems there are several different definitions of the structure at infinity; see, e.g., [17], [14]. We call the structure at infinity that was defined in [17] the *geometric* structure at infinity. This structure at infinity has proved its importance in, e.g., the solution of the static state feedback input-output decoupling problem (cf. [17]) and in obtaining (sufficient) conditions for solvability of the nonlinear model matching problem (cf. [6]). Here we will use the algebraic definition of [14], which in general yields a different structure at infinity than the geometric one. This structure at infinity has already proved its importance in the dynamic state feedback input-output decoupling problem and in obtaining sufficient conditions for solvability of the nonlinear model matching problem (cf. [16]) that are weaker than the conditions obtained in [6]. To repeat this algebraic definition, we consider Singh's algorithm as described in § 2 and define the following integers:

$$(33) \quad \begin{aligned} \pi_1 &:= \rho^*, \\ \pi_k &:= \rho^* - \rho_{k-1}, \quad k \geq 2, \end{aligned}$$

$$\nu_i := \text{the number of } \pi_k \text{'s that are greater than or equal to } i, \quad i \geq 1.$$

DEFINITION 3.1 (see [14]). The *algebraic* structure at infinity of the system Σ_0 consists of the set $\{\nu_i\}$ of orders of the zeros at infinity, together with the set $\{\pi_k\}$ of numbers of zeros at infinity whose order is greater than or equal to k .

Essentially, the sets $\{\nu_i\}$, $\{\pi_k\}$, and $\{\rho_k\}$ contain the same information about the algebraic structure at infinity. In the following, we restrict ourselves to considering the set $\{\rho_k\}$.

Let Σ_q denote the system Σ where the disturbances are considered to be an extra set of inputs. For Σ_0 and Σ_q , the sets defining their algebraic structure at infinity are denoted by $\{\rho_{0k}\}$, $\{\rho_{qk}\}$, respectively.

THEOREM 3.2. Consider the analytic system Σ and assume that it satisfies (A1), (A2). Let x_0 be a strongly regular point for Σ . Then the DDDPdm is locally solvable around x_0 if and only if Σ_q and Σ_0 have the same algebraic structure at infinity.

Proof. Consider the first step of Singh's algorithm applied to Σ_0 and Σ_q around x_0 and employ the notation of the sufficiency part of the proof of Theorem 2.6. As was shown in this proof, condition (23) is satisfied for $k=0$ if and only if there exists a $\sigma_1(x)$ such that (27) holds. Hence the fact that (23) holds for $k=0$ is equivalent to

$$(34) \quad \rho_{q1} = \text{rank} \begin{pmatrix} \tilde{b}_1 & \tilde{c}_1 \\ \hat{b}_1 & \hat{c}_1 \end{pmatrix} = \text{rank} \begin{pmatrix} \tilde{b}_1 & \tilde{b}_1 \sigma_1 \\ \hat{b}_1 & \hat{b}_1 \sigma_1 \end{pmatrix} = \text{rank} \begin{pmatrix} \tilde{b}_1 \\ \hat{b}_1 \end{pmatrix} = \rho_{01}.$$

Applying the above arguments repeatedly, we show that (23) holds for $k=0, \dots, n-1$ if and only if $\rho_{qk} = \rho_{0k}$ for $k=1, \dots, n$, which establishes our claim. \square

Remark 3.3. Recall that for linear systems the geometric and algebraic structure at infinity are the same (cf. [14]). Moreover, recall from, e.g., [22] that for linear systems the DDPdm is solvable if and only if Σ_0 and Σ_q have the same structure at infinity.

We will now use the result of Theorem 3.2 to give conditions for the local solvability of the DDDP for Σ , using an idea from [22]. To do this we introduce an auxiliary system Σ_a , which is defined as

$$(35) \quad \Sigma_a \begin{cases} \dot{x} = f(x) + g(x)w + p(x)q \\ \dot{w} = v \\ y = h(x), \end{cases}$$

where v denotes the inputs of Σ_a . Let Σ_{a0} denote the system Σ_a with $q \equiv 0$ and let Σ_{aq} denote the system Σ_a where the disturbances are considered to be an extra set of inputs.

THEOREM 3.4. *Consider the analytic system Σ and assume that it satisfies (A1), (A2). Let x_0 be a strongly regular point for Σ . Then the DDDP is locally solvable around x_0 if and only if Σ_{aq} and Σ_{a0} have the same algebraic structure at infinity.*

Proof. Necessity. Assume that the DDDP is locally solvable around x_0 by means of a compensator

$$(36) \quad R \begin{cases} \dot{z} = \alpha(x, z) + \beta(x, z)\tilde{u} \\ u = \gamma(x, z) + \delta(x, z)\tilde{u}. \end{cases}$$

Consider the following compensator for Σ_a :

$$(37) \quad R_a \begin{cases} \dot{z}_1 = \alpha(x, z_1) + \beta(x, z_1)z_2 \\ \dot{z}_2 = \hat{u} \\ v = \sigma(x, w, z_1, z_2, q), \end{cases}$$

with

$$\begin{aligned} \sigma(x, w, z_1, z_2, q) = & \left[\frac{\partial \gamma}{\partial x}(x, z_1) + \frac{\partial \delta}{\partial x}(x, z_1)z_2 \right] [f(x) + g(x)w + p(x)q] \\ & + \left[\frac{\partial \gamma}{\partial z_1}(x, z_1) + \frac{\partial \delta}{\partial z_1}(x, z_1)z_2 \right] [\alpha(x, z_1) + \beta(x, z_1)z_2] + \delta(x, z_1)\hat{u} \end{aligned}$$

and \hat{u} denoting the new inputs. Then we find that

$$w = \int v \, dt = \gamma(x, z_1) + \delta(x, z_1)z_2,$$

and thus R_a locally solves the DDDPdm for Σ_a , since R solves the DDDP for Σ . Furthermore, we have as an (almost) immediate consequence of the fact that R is invertible, that R_a is also invertible. Thus, the DDDPdm is locally solvable for Σ_a and hence by Theorem 3.2, Σ_{aq} and Σ_{a0} have the same algebraic structure at infinity.

Sufficiency. Assume that Σ_{aq} and Σ_{a0} have the same algebraic structure at infinity; i.e., the DDDPdm is locally solvable for Σ_a , say, by means of a compensator

$$(38) \quad Q_a \begin{cases} \dot{z} = \alpha(x, w, z, q) + \beta(x, w, z, q)s \\ v = \gamma(x, w, z, q) + \delta(x, w, z, q)s, \end{cases}$$

with s denoting the new inputs. Then it is obvious that the compensator

$$(39) \quad Q \begin{cases} \dot{z}_1 = \gamma(x, z_1, z_2, q) + \delta(x, z_1, z_2, q)s \\ \dot{z}_2 = \alpha(x, z_1, z_2, q) + \beta(x, z_1, z_2, q)s \\ u = z_1 \end{cases}$$

locally solves the DDDPdm for Σ . Now apply Singh's algorithm to Σ_0 , yielding a reordering $\tilde{y}_1, \dots, \tilde{y}_n$ of the outputs. Employ the notation of the proof of Theorem 2.6. Then for Σ we have in particular that

$$(40) \quad \begin{aligned} \dot{\tilde{y}}_1 &= \tilde{a}_1(x) + \tilde{b}_1(x)u + \tilde{c}_1(x)q, \\ \dot{\hat{y}}_1 &= \hat{a}_1(x) + \hat{b}_1(x)u + \hat{c}_1(x)q. \end{aligned}$$

Since Q locally solves the DDDPdm for Σ , the q -dependence in (40) must have vanished if we put u in (40) equal to the output of Q . This implies that actually $\tilde{c}_1 \equiv 0$, $\hat{c}_1 \equiv 0$, since the output of Q does not depend on q . It can be checked that this implies that (22) holds for $k=0$. Applying the above arguments repeatedly, we can show that (22) holds for $k=0, \dots, n-1$. By Theorem 2.6 this implies that the DDDP is solvable around x_0 . \square

Remark 3.5. At this point it is useful to compare our (algebraic) results on the DDDP with the results of [16] on the so-called nonlinear model matching problem (MMP). In fact, as was already shown in [6], the nonlinear MMP can be related to some kind of DDPdm. However, in contrast with the situation we consider in this paper, it turns out that the corresponding state feedback solving this DDPdm need *not* be regular. Clearly, the existence of a regular solution is, of course, a sufficient condition for solvability of the nonlinear MMP. Hence by Theorem 3.2 the coincidence of certain algebraic structures at infinity is a sufficient condition for solvability of the nonlinear MMP. This is exactly the result of [16]. As the equality of the algebraic structures at infinity is only a sufficient condition for the solvability of the MMP, and *not* a necessary condition (see [16] for a counterexample), it is clear that our solution of the DDDPdm by means of *regular* dynamic state feedback cannot be cast in the results described in the aforementioned paper.

4. Geometric conditions. In this section we give intrinsic geometric conditions for local solvability of the DDDP and the DDDPdm by translating the results of § 3. We mainly restrict our attention to the DDDP. The reasoning for the DDDPdm follows the same lines.

Consider again system Σ and assume that it satisfies (A1), (A2). Let x_0 be a strongly regular point for Σ . Furthermore, assume that the DDDP for Σ is locally solvable around x_0 . Then according to Theorem 2.6 the DDDP is locally solvable by applying a Singh compensator. This compensator applied to Σ yields $y_i^{(\delta_i)} = v_i$, $i=1, \dots, m$. Obviously, the decoupling matrix (see [12], [18]) is equal to the $m \times m$ identity matrix. Hence the decoupling matrix of the composite system is of full rank. Then it is well known that the maximal locally controlled invariant distribution in $\ker dh$ for the composite system, denoted by Δ_e^* , is given by

$$(41) \quad \Delta_e^* = \bigcap_{i=1}^m \bigcap_{k=0}^{\delta_i-1} \ker dy_i^{(k)}.$$

Obviously, Δ_e^* depends on the choice of the $\tilde{y}_k^{(k)}$'s in Singh's algorithm, so Δ_e^* is by no means uniquely defined. However, the solvability of the DDDP does not depend on the way in which Singh's algorithm is performed (see Theorem 2.6). Hence, for any distribution Δ of the form (41) generated by applying Singh's algorithm to Σ_0 , we have $\mathcal{P} \times \{0\} \subset \Delta$. Consequently, the distribution $\mathcal{P} \times \{0\}$ spanned by the extended disturbance vector fields is always contained in Δ_e^* . Note that by construction Δ_e^* is contained in $T\mathcal{X} \times \{0\} \subset T\mathcal{X} \times T\mathcal{Z}$ (with abuse of notation). However, the vector fields that span Δ_e^* may very well depend on z . Since the disturbance vector fields p only depend on

x , they are contained in the (not necessarily controlled invariant) maximal subdistribution $\tilde{\Delta}_e^*$ of Δ_e^* that contains the vector fields in Δ_e^* that only depend on x . This distribution $\tilde{\Delta}_e^*$ can be found by means of the following algorithm.

ALGORITHM 4.1.

Step 1.

$$\Delta_0 := \Delta_e^*.$$

Step k .

$$\Delta_k := \left\{ \tau \in \Delta_{k-1} \left| \left[\tau, \frac{\partial}{\partial z} \right] \in \Delta_{k-1} + \text{span} \left\{ \frac{\partial}{\partial z} \right\} \right. \right\}.$$

Here $[\tau, \partial/\partial z]$ is shorthand notation for the Lie-brackets $[\tau, \partial/\partial z_i]$ for $i = 1, \dots, \sigma$.

LEMMA 4.2. Assume that the distributions Δ_k obtained in Algorithm 4.1 have constant dimension. Then for all k , $\dim \Delta_k \leq \dim \Delta_{k-1}$ and if $\Delta_{k^*} = \Delta_{k^*-1}$, then $\Delta_k = \Delta_{k^*}$ for all $k \geq k^*$.

Proof. See [12].

Assume now that Algorithm 4.1 converges to Δ_{k^*} . Then Δ_{k^*} fulfills the condition

$$(42) \quad \left[\Delta_{k^*}, \frac{\partial}{\partial z} \right] \subset \Delta_{k^*} + \text{span} \left\{ \frac{\partial}{\partial z} \right\}.$$

As can be seen in [21], it follows from (42) that the first n components of any vector field in Δ_{k^*} do not depend on z , and since $\Delta_{k^*} \subset \Delta_e^*$ (so the last σ components of vector fields in Δ_{k^*} equal zero) the vector fields in Δ_{k^*} do not depend on z at all. Moreover, by construction, Δ_{k^*} is the largest subdistribution of Δ_e^* having this property. Hence $\Delta_{k^*} = \tilde{\Delta}_e^*$.

LEMMA 4.3. $\tilde{\Delta}_e^*$ obtained in the way described above is independent of the way we apply Singh's algorithm.

Proof. Assume we have applied Singh's algorithm in two different ways, yielding Δ_{e1}^* and Δ_{e2}^* . Assume furthermore that by applying Algorithm 4.1 to these two distributions we obtain the distributions $\tilde{\Delta}_{e1}^*$ and $\tilde{\Delta}_{e2}^*$ with $\tilde{\Delta}_{e1}^* \neq \tilde{\Delta}_{e2}^*$. This implies that there are disturbance vector fields for which the DDDP is solvable by applying Singh's algorithm in one way, but not solvable by applying Singh's algorithm in the other way. This contradicts Theorem 2.6. Hence $\tilde{\Delta}_{e1}^*$ equals $\tilde{\Delta}_{e2}^*$. \square

Let $\tilde{\Delta}^*$ be defined as the projection of the distribution $\tilde{\Delta}_e^*$ on the $T\mathcal{X}$ -space. Note that, since (42) holds, $\tilde{\Delta}^*$ is a well-defined distribution on \mathcal{X} (cf. [21]). Then $\tilde{\Delta}^*$ contains all possible vector fields that can be decoupled from the outputs by dynamic state feedback.

For the DDDPdm the reasoning is slightly different, although it follows the same lines. Here, we apply Algorithm 4.1 starting from the (not necessarily involutive) distribution $\Delta_0 = \Delta_{e\mathcal{G}}^* := \Delta_e^* + \mathcal{G} \times \{0\}$, resulting in the distribution $\tilde{\Delta}_{e\mathcal{G}}^* := \Delta_{k^*}$. Analogously to Lemma 4.3, it is then possible to prove Lemma 4.4.

LEMMA 4.4. $\tilde{\Delta}_{e\mathcal{G}}^*$ obtained in the way described above is independent of the way we apply Singh's algorithm.

Now let $\tilde{\Delta}_{\mathcal{G}}^*$ be defined as the projection of the distribution $\tilde{\Delta}_{e\mathcal{G}}^*$ on the $T\mathcal{X}$ -space. Again, this is a well-defined distribution on \mathcal{X} (cf. [21]). It is easy to see that $\tilde{\Delta}_{\mathcal{G}}^*$ contains all possible vector fields that can be decoupled from the outputs by dynamic state feedback with disturbance measurements. Note that in particular $\mathcal{G} \subset \tilde{\Delta}_{\mathcal{G}}^*$. Summarizing, we have the following theorem.

THEOREM 4.5. *Consider the system Σ and assume that it satisfies (A1), (A2). Let x_0 be a strongly regular point for Σ . Then*

1. *The DDDPdm is locally solvable around x_0 if and only if*

$$(43) \quad \mathcal{P} \subset \tilde{\Delta}_{\mathcal{G}}^*.$$

2. *The DDDP is locally solvable around x_0 if and only if*

$$(44) \quad \mathcal{P} \subset \tilde{\Delta}^*.$$

In the theorem above, we have given geometric conditions under which the DDDPdm is locally solvable. Now, of course, the question arises as to when the problem is solvable by means of a static state feedback. For this we will have to calculate Δ^* , the maximal locally controlled invariant distribution for Σ contained in $\ker dh$. Obviously, Δ^* is contained in $\tilde{\Delta}^*$, because Δ^* consists of the set of disturbance vector fields that can be decoupled by static state feedback. Observing that also $\tilde{\Delta}^*$ is contained in $\ker dh$, it is easily seen that Δ^* is the maximally locally controlled invariant distribution contained in $\tilde{\Delta}^*$. Hence Δ^* can be calculated by applying the Controlled Invariant Distribution Algorithm (cf. [12]) starting from $\tilde{\Delta}^*$.

Remark 4.6. Note that if the dimension of Δ^* equals the dimension of the zero dynamics manifold, then the DDDP is solvable if and only if the DDP is solvable.

Theorem 4.5 applied to the example of § 2 yields the following result.

Example 4.7. Consider again the system of Example 2.2. Choose $\tilde{y}_1 = y_1$. Then

$$(45) \quad \Delta_e^* = \text{span} \left\{ \frac{\partial}{\partial x_5}, x_2(x_2 + z) \frac{\partial}{\partial x_2} - (x_2^2 - x_4 z) \frac{\partial}{\partial x_4} \right\}$$

and

$$(46) \quad \tilde{\Delta}^* = \text{span} \left\{ \frac{\partial}{\partial x_5} \right\}.$$

5. Example. Consider the voltage frequency controlled induction motor that was described in [3]. As state variables we take the projections of the stator current and flux vectors on a reference frame (α, β) , which is fixed to the stator windings, and the angular position of the voltage input vector. As inputs we take the amplitude of the voltage input vector and the voltage supply frequency. The parameters R_s and R_r are the stator and rotor resistances, L_s and L_r are the stator and rotor self-inductances, and M is the mutual inductance. The speed ω can be considered as a slowly varying parameter, due to the large separation of timescales between the mechanical and the electromagnetic dynamics. In the following, we will assume it to be constant.

We define $\bar{x} = (x_1, \dots, x_4)$ and $x = (\bar{x}, x_5)$, and we assume that a one-dimensional disturbance q influences the dynamics through the disturbance vector field $p(x) = (x_3 \ x_4 \ 0 \ 0 \ 0)^T$. Then the state equations are written as

$$(47) \quad \dot{x} = \begin{pmatrix} A\bar{x} \\ 0 \end{pmatrix} + \begin{pmatrix} g_1(x_5) & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} + p(x)q,$$

where

$$(48) \quad A = \begin{pmatrix} -(\alpha + \beta) & -\omega & \beta/L_s & \omega/\sigma L_s \\ \omega & -(\alpha + \beta) & -\omega/\sigma L_s & \beta/L_s \\ -\alpha\sigma L_s & 0 & 0 & 0 \\ 0 & -\alpha\sigma L_s & 0 & 0 \end{pmatrix}, \quad g_1(x_5) = \begin{pmatrix} \cos x_5/\sigma L_s \\ \sin x_5/\sigma L_s \\ \cos x_5 \\ \sin x_5 \end{pmatrix},$$

and $\alpha = R_s/\sigma L_s$, $\beta = R_r/\sigma L_r$, $\sigma = 1 - (M^2/L_s L_r)$.

Suitable outputs for the system are defined in terms of the stator flux and the torque. Hence, the following nonlinear output functions will be used:

$$(49) \quad \begin{aligned} h_1(x) &= \Phi_s^2 = x_3^2 + x_4^2, \\ h_2(x) &= T_m = x_2x_3 - x_1x_4. \end{aligned}$$

Applying the Controlled Invariant Distribution Algorithm (cf. [12]), using REDUCE, we find that $\Delta^* = \{0\}$. Hence neither DDPm nor DDP is solvable for (47), (49). However, by applying Singh's algorithm to (47), (49), we find that we can solve the DDDPm by applying the following Singh compensator with disturbance feedthrough:

$$(50) \quad \begin{aligned} \dot{z} &= v_1, \\ u_1 &= \phi_1(x, z), \\ u_2 &= \phi_2(x, z, q, v_1, v_2), \end{aligned}$$

where

$$(51) \quad \phi_1(x, z) = \frac{2\alpha\sigma L_s(x_1x_3 + x_2x_4) + z}{2(x_3 \cos x_5 + x_4 \sin x_5)},$$

and where $\phi_2(x, z, q, v_1, v_2)$ can be calculated from

$$(52) \quad \begin{aligned} \phi_2(x, z, q, v_1, v_2) &= \frac{1}{\mathcal{L}_{g_2}\mathcal{L}_f y_2 + \phi_1 \mathcal{L}_{g_2}\mathcal{L}_{g_1} y_2} \left[v_2 - \frac{\partial \phi_1}{\partial z} \mathcal{L}_{g_1} y_2 \cdot v_1 \right. \\ &\quad - (\mathcal{L}_{f+\phi_1 g_1} \mathcal{L}_f y_2 + \mathcal{L}_{g_1} y_2 \mathcal{L}_{f+\phi_1 g_1} \phi_1 + \phi_1 \mathcal{L}_{f+\phi_1 g_1} \mathcal{L}_{g_1} y_2) \\ &\quad \left. - q(\mathcal{L}_p \mathcal{L}_f y_2 + \mathcal{L}_{g_1} y_2 \mathcal{L}_p \phi_1 + \phi_1 \mathcal{L}_p \mathcal{L}_{g_1} y_2) \right]. \end{aligned}$$

6. Conclusions. In this paper we have introduced the dynamic disturbance decoupling problem (DDDP) for nonlinear systems, and a local solution is obtained in the case where the system under consideration is invertible. The solution is given in both an algebraic and a (differential) geometric way. This clearly exhibits that the DDDP forms a proper extension of the standard (static) DDP (cf. [13]). As stated our solution is obtained for invertible (square) nonlinear systems. An extension of the results in this paper to nonsquare, noninvertible systems can be found in [10].

Acknowledgments. The authors would like to thank A. J. van der Schaft for some very useful discussions and suggestions. The second author gratefully acknowledges some very useful preliminary discussions with Prof. U. Kotta on the topic of this paper. Finally, we have learned that very recently W. Respondek (Warsaw) has obtained similar results on the DDDP using a different type of dynamic compensator than ours.

REFERENCES

- [1] G. BASILE AND G. MARRO, *Controlled and conditioned invariant subspaces in linear system theory*, J. Optim. Theory Appl., 3 (1969), pp. 306–315.
- [2] S. P. BHATTACHARYYA, *Disturbance rejection in linear systems*, Internat. J. Control, 5 (1974), pp. 633–637.
- [3] A. DE LUCA AND G. ULIVI, *Dynamic decoupling of voltage frequency controlled induction motors*, in Analysis and Optimization of Systems, A. Bensoussan and J. L. Lions, eds., Lecture Notes in Control and Inform. Sci., 111, Springer, Berlin, 1988, pp. 127–137.

- [4] M. D. DI BENEDETTO AND J. W. GRIZZLE, *Intrinsic notions of regularity for local inversion, output nulling, and dynamic extension of nonsquare systems*, Control Theory Adv. Tech., 6 (1990), pp. 357–381.
- [5] M. D. DI BENEDETTO, J. W. GRIZZLE, AND C. H. MOOG, *Rank invariants of nonlinear systems*, SIAM J. Control Optim., 27 (1989), pp. 658–672.
- [6] M. D. DI BENEDETTO AND A. ISIDORI, *The matching of nonlinear models via dynamic state feedback*, SIAM J. Control Optim., 24 (1986), pp. 1063–1075.
- [7] A. GLUMINEAU AND C. H. MOOG, *The essential orders and the nonlinear decoupling problem*, Internat. J. Control, 50 (1989), pp. 1825–1834.
- [8] R. M. HIRSCHORN, *Invertibility of multivariable nonlinear control systems*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 855–865.
- [9] ———, *(A, B)-invariant distributions and disturbance decoupling of nonlinear systems*, SIAM J. Control Optim., 19 (1981), pp. 1–19.
- [10] H. J. C. HUIJBERTS, H. NIJMEIJER, AND L. L. M. VAN DER WEGEN, *Dynamic disturbance decoupling for nonlinear systems: The nonsquare and noninvertible case*, in Controlled Dynamical Systems, B. Bonnard, B. Bride, J. P. Gauthier, and I. Kupka, eds., Birkhäuser, Boston, 1990, to appear.
- [11] ———, *Minimality of dynamic input-output decoupling for nonlinear systems*, preprint, submitted for publication.
- [12] A. ISIDORI, *Nonlinear Control Systems*, 2nd ed., Springer, Berlin, 1989.
- [13] A. ISIDORI, A. J. KRENER, C. GORI-GIORGI, AND S. MONACO, *Nonlinear decoupling via feedback: A differential geometric approach*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 331–345.
- [14] C. H. MOOG, *Nonlinear decoupling and structure at infinity*, Math. Control Signals Systems, 1 (1988), pp. 257–268.
- [15] ———, *Note on the left-invertibility of nonlinear systems*, in Analysis and Control of Nonlinear Systems, C. I. Byrnes, C. F. Martin, and R. E. Sacks, eds., Elsevier, Amsterdam, 1988, pp. 469–474.
- [16] C. H. MOOG, A. M. PERDON, AND G. CONTE, *Model matching and factorization for nonlinear systems: A structural approach*, SIAM J. Control Optim., 29 (1990), pp. 769–785.
- [17] H. NIJMEIJER AND J. M. SCHUMACHER, *Zeros at infinity for affine nonlinear control systems*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 566–573.
- [18] H. NIJMEIJER AND A. J. VAN DER SCHAFT, *Nonlinear Dynamical Control Systems*, Springer, New York, 1990.
- [19] S. N. SINGH, *A modified algorithm for invertibility in nonlinear systems*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 595–598.
- [20] ———, *Generalised decoupled-control synthesis for invertible nonlinear systems*, IEE Proceedings, Vol. 128, Pt.D., 1981, pp. 157–161.
- [21] A. J. VAN DER SCHAFT, *Observability and controllability for smooth nonlinear systems*, SIAM J. Control Optim., 20 (1982), pp. 338–354.
- [22] J. W. VAN DER WOUDE, *On the structure at infinity of a structured system*, Report BS-R89, Centre for Mathematics and Computer Science, Amsterdam, the Netherlands, 1989.
- [23] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, 3rd ed., Springer, New York, 1985.

A NONREGULAR SOLUTION OF THE NONLINEAR DYNAMIC DISTURBANCE DECOUPLING PROBLEM WITH AN APPLICATION TO A COMPLETE SOLUTION OF THE NONLINEAR MODEL MATCHING PROBLEM*

H. J. C. HUIJBERTS†

Abstract. The nonregular dynamic disturbance decoupling problem for nonlinear control systems is introduced. A local solution is given by means of a constructive algorithm that is based on Singh's algorithm and the clamped dynamics algorithm. Further studied is the nonlinear model matching problem that is defined as follows: given a nonlinear control system, to be referred to as the plant, and another nonlinear control system, to be referred to as the model, can a compensator for the plant be found in such a way that the input-output behavior of the compensated plant matches that of the model? By proving that the solvability of the nonlinear model matching problem is equivalent to the solvability of an associated nonregular dynamic disturbance decoupling problem, a complete local solution of this problem can be established.

Key words. nonlinear control systems, dynamic disturbance decoupling, dynamic precompensation, clamped dynamics algorithm, nonlinear model matching

AMS(MOS) subject classifications. 93C10, 93B50, 93C35

1. Introduction. Consider a nonlinear multi-input multi-output control system with disturbances, of the form

$$(1) \quad \begin{aligned} \dot{x} &= f(x) + g(x)u + p(x)q, \\ y &= h(x), \end{aligned}$$

where $x \in \mathcal{X}$, an open subset of \mathbb{R}^n , the inputs $u \in \mathbb{R}^m$, the outputs $y \in \mathbb{R}^p$, the disturbances $q \in \mathbb{R}^r$, f and h are vector-valued analytic functions, and g and p are matrix-valued analytic functions, all of appropriate dimensions. In the disturbance decoupling problem (DDP) for the system (1) we search for a regular static state feedback

$$(2) \quad u = \alpha(x) + \beta(x)v,$$

with v a new m -dimensional control and $\beta(x)$ a nonsingular $m \times m$ matrix for all x , so that in the feedback-modified dynamics

$$(3) \quad \dot{x} = f(x) + g(x)\alpha(x) + g(x)\beta(x)v + p(x)q$$

the disturbances q do not affect the outputs y . A local solution of the DDP using differential geometric tools was initiated in [21] and [14] and has led to a more or less complete understanding of this problem; see, e.g., [20], [33]. The nonlinear DDP forms a direct generalization of the linear DDP and the theory about the nonlinear DDP typically extends the well-known linear geometric theory (see [36]) to a nonlinear context.

The purpose of this paper is to study, as in [17], a dynamic version of the DDP for the nonlinear system (1). That is, instead of a static feedback law (2) we allow for a dynamic state feedback

$$(4) \quad \begin{aligned} \dot{z} &= \alpha(x, z) + \beta(x, z)v, \\ u &= \gamma(x, z) + \delta(x, z)v, \end{aligned}$$

* Received by the editors June 4, 1990; accepted for publication (in revised form) January 17, 1991.

† Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE Enschede, the Netherlands. Present address, Department of Mathematics and Computing Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, the Netherlands.

with z the μ -dimensional compensator state, and v an m -dimensional new control. In the dynamic version of the dynamic disturbance decoupling problem we require that in the modified dynamics

$$(5) \quad \begin{aligned} \dot{x} &= f(x) + g(x)\gamma(x, z) + g(x)\delta(x, z)v + p(x)q, \\ \dot{z} &= \alpha(x, z) + \beta(x, z)v \end{aligned}$$

the disturbances q do not influence the outputs y . In [17] a *regular* version of this problem was studied, where regularity means that we demand the system (4) with inputs v and outputs u to be invertible for all x . We refer to this problem as the dynamic disturbance decoupling problem (DDDP). In this paper we solve a version of the DDDP where we drop the requirement of regularity of the compensator (4). This problem will be referred to as the nonregular dynamic disturbance decoupling problem (nDDDP). The reason for studying this problem is that if the DDDP is not solvable, we may still be able to solve the nDDDP, albeit at the expense of some of the effective controls. The solution is given by means of an algorithm based on the clamped dynamics algorithm (cf. [22], [35]) and Singh's algorithm (cf. [34], [9]). Basically, the method used in this paper to solve the nDDDP is an extension of the method given in [17] to solve the DDDP. Here the basic tool was Singh's algorithm. One special feature of the algorithm is that it is constructive. A drawback is that it does not give a compensator of minimal order. Therefore we cannot make any statement about the solvability of the nonregular disturbance decoupling problem by means of static state feedback.

The solution of the nDDDP presented in this paper also turns out to be of use for solving another synthesis problem: the nonlinear model matching problem (MMP). This problem can be formulated as follows: Given an affine nonlinear control system, to be referred to as the plant P , and another affine nonlinear control system, to be referred to as the model M , can we find a compensator for P such that the input-output behavior of the precompensated plant matches that of M ?

For linear systems this problem is completely solved (see, e.g., [11], [27], [30]–[32]). For nonlinear systems until now only partial solutions have appeared (see, e.g., [6], [7], [10], [15], [18], [19], [29]). It will be shown in this paper that the nonlinear MMP can be formulated as an nDDDP with disturbance measurements. This observation was already made for linear systems in [32], [11]. In [10] sufficient conditions for solvability of the nonlinear MMP were given by solving an associated DDP with disturbance measurements by means of regular static state feedback. Furthermore, the sufficient conditions for solvability of the nonlinear MMP that were given in [29] can be viewed as following from the solution of an associated DDDP with disturbance measurements by means of regular dynamic state feedback as was also studied in [17]. It is important to note that the partial solutions to the MMP mentioned above are all given in terms of structural invariants of the system under consideration, whereas the complete solution given in this paper is in terms of an algorithm.

The paper is organized as follows. In § 2 we give the clamped dynamics algorithm. This algorithm allows us to determine the clamped dynamics of a nonlinear control system, i.e., the internal dynamics that are compatible with the constraint that the output is zero for all times. Furthermore, we give an important proposition that gives a connection between the clamped dynamics of a nonlinear control system and the clamped dynamics of the same system precompensated by a dynamic state feedback. In § 3 we introduce the nonregular dynamic disturbance decoupling problem with disturbance measurements (nDDDPdm) and the nDDDP, and we present an algorithm for solving both problems. In § 4 we formulate the nonlinear MMP and give a complete

local solution to this problem by associating an nDDDPdm with it. In § 5, finally, some conclusions will be drawn.

2. The clamped dynamics algorithm. In this section we introduce the notion of *clamped dynamics* and an algorithm to compute these, following [35]. Consider a nonlinear control system of the form

$$(6) \quad \begin{aligned} \dot{x} &= f(x) + g(x)u, \\ y &= h(x), \end{aligned}$$

where $x \in \mathcal{X}$, an open subset of \mathbb{R}^n , the inputs $u \in \mathbb{R}^m$, the outputs $y \in \mathbb{R}^p$, f and h are vector-valued analytic functions, and g is a matrix-valued analytic function, all of appropriate dimensions. The *clamped dynamics* of (6) are defined as the internal dynamics of (6) that are compatible with the constraint that the output is zero for all times. The notion of clamped dynamics was first identified, in the single-input single-output case, in [4], [28] (note, however, that in these references the notion was called zero dynamics). For the multi-input multi-output case it was further elaborated in [5], [22], [35]. In [22], a general algorithm to calculate the clamped dynamics was proposed. This algorithm was based on a modified version of Hirschorn's structure algorithm (see [13]). A different (but equivalent) algorithm was proposed in [35] and is based upon a modified version of Krener's algorithm (see [25]). In this paper we follow [35].

DEFINITION 2.1. A submanifold $N \subset \mathcal{X}$ is *controlled invariant* for (6) if there exists an analytic feedback $u = \alpha(x)$ such that the vector field $f(x) + g(x)\alpha(x)$ is tangent to N .

Remark 2.2. If $\dim(TN + \mathcal{G}) = \text{constant}$, where \mathcal{G} is the distribution spanned by the columns of g , this definition is equivalent to the following characterization: N is controlled invariant if for any $x_0 \in N$ there exists an admissible control $\bar{u}(t)$ such that the solution of $\dot{x}(t) = f(x) + g(x)\bar{u}(t)$, $x(0) = x_0$ remains in N (see [33]).

DEFINITION 2.3. $N \subset \mathcal{X}$ is an *output-nulling* controlled invariant submanifold for (6) if there exists a feedback $u = \alpha(x)$, such that $f(x) + g(x)\alpha(x)$ is tangent to N and $h(x)$ is zero on N .

To determine the dynamics compatible with the constraints $y(t) = 0$ for all t , we need to compute the maximal output-nulling controlled invariant submanifold, provided it exists. This can be done by means of the following algorithm that is based on the algorithm from [35].

ALGORITHM 2.4. Clamped dynamics algorithm. Consider the system (6). Let $x_0 \in \mathcal{X}$ be such that $h(x_0) = 0$ and assume that $h = (h_1, \dots, h_p)$ has constant rank p in a neighborhood of x_0 in $h^{-1}(x_0)$.

- Step 0

Locally around x_0 the set $N_0 = h^{-1}(0)$ is an $(n - p_0)$ -dimensional submanifold, where $p_0 := p$. Denote $\phi_0 := h$.

- Step k

Let N_{k-1} be a smooth $(n - p_{k-1})$ -dimensional manifold through x_0 , given as $N_{k-1} = \{x \mid \phi_{k-1}(x) = 0\}$. Calculate

$$(7) \quad \dot{\phi}_{k-1} = \frac{\partial \phi_{k-1}}{\partial x} [f(x) + g(x)u] =: A_k(x) + B_k(x)u.$$

Assume that $B_k(x)$ has constant rank r_k in a neighborhood of x_0 in N_{k-1} . After a possible permutation of the entries of ϕ_{k-1} we may assume that the first r_k

rows of B_k are linearly independent. Accordingly, we write (7) as

$$(8) \quad \begin{pmatrix} \dot{\phi}_{\bar{k}} \\ \dot{\phi}_{\bar{\ell}} \end{pmatrix} = \begin{pmatrix} \tilde{A}_k(x) + \tilde{B}_k(x)u \\ \hat{A}_k(x) + \hat{B}_k(x)u \end{pmatrix},$$

where $\tilde{B}_k(x)$ has full row rank r_k in a neighborhood of x_0 in N_{k-1} . Let $\tilde{B}_k^+(x)$ be a right inverse of $\tilde{B}_k(x)$. Letting $u = -\tilde{B}_k^+(x)\tilde{A}_k(x)$, we find from (8) that

$$(9) \quad \begin{pmatrix} \dot{\phi}_{\bar{k}} \\ \dot{\phi}_{\bar{\ell}} \end{pmatrix} = \begin{pmatrix} 0 \\ \hat{A}_k(x) - \hat{B}_k(x)\tilde{B}_k^+(x)\tilde{A}_k(x) \end{pmatrix} = \begin{pmatrix} 0 \\ \bar{\phi}_k(x) \end{pmatrix}.$$

Note that, since each row of $\hat{B}_k(x)$ is linearly dependent on the rows of $\tilde{B}_k(x)$, $\bar{\phi}_k(x)$ is independent of the choice of $\tilde{B}_k^+(x)$. Assume that $\bar{\phi}_k(x_0) = 0$ and that $\bar{\phi}_k(x)$ has constant rank s_k in a neighborhood of x_0 in N_{k-1} . Then locally around x_0 , $N_k := \{x \in N_{k-1} \mid \bar{\phi}_k(x) = 0\}$ is an $(n - p_k)$ -dimensional submanifold, with $p_k := p_{k-1} + s_k$. Permute the entries of $\bar{\phi}_k$ such that the first s_k entries are independent on N_k , and denote $\phi_k := (\phi_{k-1}, \phi_{k1}, \dots, \phi_{ks_k})$.

If, at every step of the algorithm, the two constant rank assumptions are satisfied and $\bar{\phi}_k(x_0) = 0$, then we call x_0 a *regular point* for the algorithm. If x_0 is a regular point for the clamped dynamics algorithm, then it easily follows that the algorithm terminates after $k^* < n$ iterations, where k^* is the least integer such that $N_{k^*-1} = N_{k^*}$, or equivalently k^* is the least integer for which $\bar{\phi}_k \equiv 0$ on N_{k-1} . We will call the maximal connected component of N_{k^*} containing x_0 the *clamped dynamics manifold* of (6) and we denote it by N^* . A control that renders the dynamics on N^* invariant is given by $u = -\tilde{B}_{k^*}^+(x)\tilde{A}_{k^*}(x)$.

In what follows we will control nonlinear systems of the form (6) by means of a compensator R of the form

$$(10) \quad R \begin{cases} \dot{z} = \alpha(x, z) + \beta(x, z)v \\ u = \gamma(x, z) + \delta(x, z)v \end{cases}$$

with $z \in \mathbb{R}^\mu$, v denoting the new inputs and real analytic $\alpha, \beta, \gamma, \delta$. A question that arises is: what is the connection between the clamped dynamics manifold of (6) and the clamped dynamics manifold of (6), (10)? The following proposition gives an answer to this question.

PROPOSITION 2.5. *Let the clamped dynamics manifold of (6) be given by $N^* = \{x \mid \phi_{k^*}(x) = 0\}$. Then the clamped dynamics manifold of (6), (10) is given by $M^* = \{(x, z) \mid \phi_{k^*}(x) = 0, \psi(x, z) = 0\}$ for some vector of functions $\psi(x, z)$.*

The proof appears in the Appendix.

3. The nonregular dynamic disturbance decoupling problem (nDDDP). In this section we formulate two kinds of nDDDPs and give a local solution for both problems.

Let Σ be a nonlinear multi-input multi-output control system with disturbances of the form

$$(11) \quad \Sigma \begin{cases} \dot{x} = f(x) + g(x)u + p(x)q \\ y = h(x), \end{cases}$$

where $x \in \mathcal{X}$, an open subset of \mathbb{R}^n , the inputs $u \in \mathbb{R}^m$, the outputs $y \in \mathbb{R}^p$, the disturbances $q \in \mathbb{R}^r$, f and h are vector-valued analytic functions, and g and p are matrix-valued analytic functions, all of appropriate dimensions. Analogously to [17] we define the following problems for Σ .

DEFINITION 3.6. Consider the system Σ and let a point $x_0 \in \mathcal{X}$ be given.

1. The local nDDDP with disturbance measurements (nDDDPdm) consists in finding (if possible) a compensator Q of the form

$$(12) \quad Q \begin{cases} \dot{z} = \alpha(x, q, z) + \beta(x, q, z)v \\ u = \gamma(x, q, z) + \delta(x, q, z)v \end{cases}$$

with $z \in \mathbb{R}^\mu$, a neighborhood U of x_0 in \mathcal{X} , an open subset V of \mathbb{R}^μ , and a map $F: U \rightarrow V$ with the property that $y^{\Sigma \circ Q}(\bar{x}, F(\bar{x}), t)$ is independent of q for all $\bar{x} \in U$. Here $y^{\Sigma \circ Q}(\bar{x}, F(\bar{x}), t)$ denotes the output of the precompensated system $\Sigma \circ Q$ initialized at $(\bar{x}, F(\bar{x}))$, at time t .

2. The local nDDDP consists in finding (if possible) a compensator R of the form

$$(13) \quad R \begin{cases} \dot{z} = \alpha(x, z) + \beta(x, z)v \\ u = \gamma(x, z) + \delta(x, z)v \end{cases}$$

with $z \in \mathbb{R}^\mu$, a neighborhood U of x_0 in \mathcal{X} , an open subset V of \mathbb{R}^μ , and a map $F: U \rightarrow V$ with the property that $y^{\Sigma \circ Q}(\bar{x}, F(\bar{x}), t)$ is independent of q for all $\bar{x} \in U$.

If we require the compensators Q and R to be invertible, the above problems will be referred to by DDDPdm and DDDP, respectively.

A solution of the DDDP and the DDDPdm can be found in [17]. Recall that for linear systems the conditions for solvability of nDDDPdm, DDPdm, nDDDPdm, and DDDPdm are all equivalent (see, e.g., [1], [2]). From Example 3.2 in [17] it has become clear that for nonlinear systems solvability of the DDP is not equivalent to the solvability of the DDDP. The following example illustrates that for nonlinear systems solvability of the nDDDP is not equivalent to solvability of the DDP nor the DDDP.

Example 3.7. Consider the nonlinear system

$$(14) \quad \begin{aligned} \dot{x}_1 &= x_2 u, \\ \dot{x}_2 &= x_3 + q, \\ \dot{x}_3 &= x_1 u, \\ y_1 &= x_1, \\ y_2 &= x_3. \end{aligned}$$

It is easy to see that we can solve the nDDDP for this system by putting $u = 0$, while using the theory developed in [17] it can be shown that the DDDP is not solvable for (14), which also implies that the DDP is not solvable for (14).

In what follows we will give a constructive algorithm to solve the nDDDPdm. In order to get a better insight into how the algorithm works, we first discuss the ideas behind the algorithm. So, consider a system Σ with disturbances, of the form (11). Choose a point $x_0 \in \mathcal{X}$. First assume that we want to solve the nDDDPdm around x_0 . Clearly, a compensator Q solves the nDDDPdm for Σ around x_0 if and only if for $\Sigma \circ Q$ the $y_i^{(j)}(t)$ do not depend on q for all t . We now apply the first steps of Singh's algorithm (cf. [34], [9]) to Σ . Thus, for Σ we determine

$$(15) \quad \dot{y} = \frac{\partial h}{\partial x}(x)[f(x) + g(x)u + p(x)q] =: a_1(x) + b_1(x)u + c_1(x)q.$$

Assume that $b_1(x)$ has constant rank ρ_1 in a neighborhood of x_0 . After a possible permutation of the entries of y , we may assume that the first ρ_1 rows of $b_1(x)$ are linearly independent. Accordingly, we write (15) as

$$(16) \quad \begin{pmatrix} \dot{\hat{y}}_1 \\ \hat{y}_1 \end{pmatrix} = \begin{pmatrix} \tilde{a}_1(x) + \tilde{c}_1(x)q \\ \hat{a}_1(x) + \hat{c}_1(x)q \end{pmatrix} + \begin{pmatrix} \tilde{b}_1(x) \\ \hat{b}_1(x) \end{pmatrix} u,$$

where $\tilde{b}_1(x)$ has full row rank ρ_1 in a neighborhood of x_0 . Let $\tilde{b}_1^+(x)$ be a right inverse of $\tilde{b}_1(x)$. Then from the upper part of (16) it follows that

$$(17) \quad u = \tilde{b}_1^+(x)(\dot{\hat{y}}_1 - \tilde{a}_1(x) - \tilde{c}_1(x)q).$$

Combining (16) and (17) we obtain

$$(18) \quad \dot{\hat{y}}_1 = \hat{a}_1(x) + \hat{b}_1(x)\tilde{b}_1^+(x)(\dot{\hat{y}}_1 - \tilde{a}_1(x) - \tilde{c}_1(x)q) + (\hat{c}_1(x) - \hat{b}_1(x)\tilde{b}_1^+(x)\tilde{c}_1(x))q =: \psi_1(x, q, \dot{\hat{y}}_1).$$

Note that ψ_1 is affine in q and $\dot{\hat{y}}_1$. Define the matrix-valued function

$$(19) \quad D_1(x) := \frac{\partial \psi_1(x, q, \dot{\hat{y}}_1)}{\partial q}.$$

Assume that $D_1 \neq 0$. This implies that $\dot{\hat{y}}_1$ explicitly depends on q . Furthermore, this dependence is intrinsic, meaning that with another partition (\tilde{y}_1, \hat{y}_1) of y also a q -dependence would occur via a matrix-valued function $\tilde{D}_1(x)$ with the property that $\tilde{D}_1(x) = T_1(x)D_1(x)$, where $T_1(x)$ is invertible (cf. [9]). Using similar arguments as in the proof of Proposition 2.5, we can show that this q -dependence can only be resolved by means of dynamic compensation if we can choose the compensator and its initial conditions in such a way that $D_1(x(t)) = 0$ for all t . However, this would imply that the nDDDPdm is not solvable in a neighborhood of x_0 in \mathcal{X} , but at most in a neighborhood of x_0 in $\{x \mid D_1(x) = 0\}$. Thus we must necessarily have that $D_1 \equiv 0$ if we want to solve the nDDDPdm. Assuming that $D_1 \equiv 0$ we proceed by determining

$$(20) \quad \begin{aligned} \hat{y}_1^{(2)} &= \frac{\partial \psi_1}{\partial x}(x, \dot{\hat{y}}_1)[f(x) + g(x)u + p(x)q] + \frac{\partial \psi_1}{\partial \dot{\hat{y}}_1}(x, \dot{\hat{y}}_1)\dot{\hat{y}}_1^{(2)} \\ &=: a_2(x, \dot{\hat{y}}_1, \tilde{y}_1^{(2)}) + b_2(x, \dot{\hat{y}}_1)u + c_2(x, \dot{\hat{y}}_1)q. \end{aligned}$$

Define $B_2 := (\tilde{b}_1^T b_2^T)^T$ and assume that $B_2(x, \dot{\hat{y}}_1)$ has constant rank ρ_2 in a neighborhood of x_0 . Then, after a possible permutation of the entries of \hat{y}_1 , we may assume that the first $\rho_2 - \rho_1$ rows of $b_2(x, \dot{\hat{y}}_1)$ are linearly independent. Accordingly, we write (20) and the upper part of (16) as

$$(21) \quad \begin{pmatrix} \dot{\hat{y}}_1 \\ \hat{y}_1^{(2)} \\ \hat{y}_2^{(2)} \end{pmatrix} = \begin{pmatrix} \tilde{a}_1(x) + \tilde{c}_1(x)q \\ \tilde{a}_2(x, \dot{\hat{y}}_1, \tilde{y}_1^{(2)}) + \tilde{c}_2(x, \dot{\hat{y}}_1)q \\ \hat{a}_2(x, \dot{\hat{y}}_1, \tilde{y}_1^{(2)}) + \hat{c}_2(x, \dot{\hat{y}}_1)q \end{pmatrix} + \begin{pmatrix} \tilde{b}_1(x) \\ \tilde{b}_2(x, \dot{\hat{y}}_1) \\ \hat{b}_2(x, \dot{\hat{y}}_1) \end{pmatrix} u,$$

where $\tilde{B}_2(x, \dot{\hat{y}}_1) := (\tilde{b}_1^T(x)\tilde{b}_2^T(x, \dot{\hat{y}}_1))^T$ has full row rank ρ_2 . Let $\tilde{B}_2^+(x, \dot{\hat{y}}_1)$ be a right inverse of $\tilde{B}_2(x, \dot{\hat{y}}_1)$ and define $\tilde{Y}_2 := (\dot{\hat{y}}_1^T \hat{y}_2^{(2)T})^T$, $\tilde{A}_2(x, \dot{\hat{y}}_1, \tilde{y}_1^{(2)}) := (\tilde{a}_1^T(x)\tilde{a}_2^T(x, \dot{\hat{y}}_1, \tilde{y}_1^{(2)}))^T$, $\tilde{C}_2(x, \dot{\hat{y}}_1) := (\tilde{c}_1^T(x)\tilde{c}_2^T(x, \dot{\hat{y}}_1))^T$. Then again we can rewrite $\hat{y}_2^{(2)}$ as

$$(22) \quad \begin{aligned} \hat{y}_2^{(2)} &= \hat{a}_2(x, \dot{\hat{y}}_1, \tilde{y}_1^{(2)}) - \hat{b}_2(x, \dot{\hat{y}}_1)\tilde{B}_2^+(x, \dot{\hat{y}}_1)(\tilde{A}_2(x, \dot{\hat{y}}_1, \tilde{y}_1^{(2)}) - \tilde{Y}_2) \\ &\quad + (\hat{c}_2(x, \dot{\hat{y}}_1) - \hat{b}_2(x, \dot{\hat{y}}_1)\tilde{B}_2^+(x, \dot{\hat{y}}_1)\tilde{C}_2(x, \dot{\hat{y}}_1))q =: \psi_2(x, q, \tilde{Y}_2). \end{aligned}$$

Define as before the matrix-valued function

$$(23) \quad D_2(x, \dot{y}_1) := \frac{\partial \psi_2(x, q, \tilde{Y}_2)}{\partial q}.$$

First assume that $D_2 \neq 0$. Then $\hat{y}_2^{(2)}$ explicitly depends on q . Again, this dependence is intrinsic in a similar sense as described above, and it can only be resolved by means of dynamic compensation if we can choose the compensator and its initial conditions in such a way that $D_2(x(t), \dot{y}_1(t)) = 0$ for all t . Now note that D_2 is a function of x and \dot{y}_1 . Thus it may be possible to find a neighborhood U of x_0 in \mathcal{X} such that for all $x \in U$ there is a $\dot{y}_1 \in \mathbb{R}^{p_1}$ such that $D_2(x, \dot{y}_1) = 0$. By the implicit function theorem, this is the case if and only if

$$(24) \quad \text{rank} \frac{\partial \hat{D}_2}{\partial \dot{y}_1} = \text{rank} \left(\frac{\partial \hat{D}_2}{\partial x} \frac{\partial \hat{D}_2}{\partial \dot{y}_1} \right),$$

where $\hat{D}_2(x, \dot{y}_1)$ is the vector of functions consisting of the nonzero entries of D_2 . If (24) holds, we can locally treat x as a “free variable,” i.e., there is a partition $(\dot{y}_{11}, \dot{y}_{12})$ of \dot{y}_1 and a function $\bar{D}_2(x, \dot{y}_{12})$ such that locally

$$(25) \quad \hat{D}_2(x, \dot{y}_1) = 0 \Leftrightarrow \dot{y}_{11} = \bar{D}_2(x, \dot{y}_{12}).$$

We now proceed by repeating the above procedure under the constraint that D_2 and its time derivatives are zero. If again an explicit q -dependence of the type described above or an explicit q -dependence of a time derivative of D_2 occurs, we add the function characterizing the dependence to the set of constraint functions and start all over again with the new set of constraint functions.

If, on the other hand, we have that $D_2 \equiv 0$, we continue with subsequent steps of Singh's algorithm until an intrinsic q -dependence occurs and then go through the whole procedure as described above. While going through the procedure, we should apply two intermediate checks. First, it may occur at a certain step that the submanifold on which the constraints are satisfied is empty. In this case the nDDDPdm is not solvable, and we can stop the procedure. Second, assume that at a certain step the set of constraint functions is given by $\Psi_k(x, \tilde{Y}_k)$. Then, to guarantee solvability of the nDDDPdm on a neighborhood of x_0 in \mathcal{X} , we must necessarily have that there is a neighborhood U of x_0 in \mathcal{X} such that for every $x \in U$ there is a \tilde{Y}_k satisfying $\Psi_k(x, \tilde{Y}_k) = 0$. By the implicit function theorem, this is the case if and only if

$$(26) \quad \text{rank} \frac{\partial \Psi_k}{\partial \tilde{Y}_k} = \text{rank} \left(\frac{\partial \Psi_k}{\partial x} \frac{\partial \Psi_k}{\partial \tilde{Y}_k} \right),$$

so that again, locally, x can be treated as a free variable. Thus, if (26) does not hold, the nDDDPdm is not solvable, and we can stop the procedure.

If we want to solve the nDDDP instead of the nDDDPdm, the procedure described above follows the same lines, with a few minor differences. Namely, while solving the nDDDPdm, at every step k we may resolve the intrinsic dependence of $\tilde{y}_1, \dots, \tilde{y}_k$ on q by choosing an appropriate input u depending on x, q and $\tilde{y}_1, \dots, \tilde{y}_k$ and its time derivatives (see (16), (17), (21)). However, if we want to solve the nDDDP, we can only resolve the intrinsic q -dependence of $\tilde{y}_1, \dots, \tilde{y}_k$ if the functions characterizing the dependence are constrained to be zero for all t , since u is not allowed to depend on q . For the procedure this means that, while solving the nDDDP, we should define $D_1(x) = (\partial \hat{y} / \partial q)(x)$, $D_2(x, \dot{y}_1) = \partial \hat{y}_1 / \partial q$, etc.

Summarizing, the algorithm we are going to present will consist of applying steps of Singh's algorithm and the clamped dynamics algorithm to Σ , together with some intermediate checks. The use of Singh's algorithm can be circumvented by observing that applying Singh's algorithm to Σ is equivalent to applying the clamped dynamics algorithm to the augmented system (cf. [35])

$$(27) \quad \Sigma_{a0} \begin{cases} \dot{x} = f(x) + g(x)u + p(x)q \\ \dot{w}_{i1} = w_{i2} \\ \dot{w}_{i2} = w_{i3} \\ \vdots \\ \dot{w}_{i\nu_0} = v_i \\ y_{0i} = h_i(x) - w_{i1} \end{cases} \quad (i = 1, \dots, p)$$

with $\nu_0 := n$. Thus we only have to use the clamped dynamics algorithm.

This leads to the following algorithm.

ALGORITHM 3.8.

• Step 0

Define the system Σ_{a0} as in (27).

• Step $k+1$

Let $\Sigma_{a0}, \dots, \Sigma_{ak}$ and ν_0, \dots, ν_k be defined. If we apply the clamped dynamics algorithm to Σ_{ak} , where we consider q as a parameter, we obtain (vectors of) functions $\phi_1^k(x, q, w), \dots, \phi_{\pi_k}^k(x, q, w)$, where π_k denotes the final step of the clamped dynamics algorithm applied to Σ_{ak} . While solving the nDDDPm, let τ_k be the smallest integer for which

$$(28) \quad 0 \neq D_{\tau_k}^k(x, w) := \frac{\partial \phi_{\tau_k+1}^k}{\partial q}.$$

While solving the nDDDP, let τ_k be the smallest integer for which

$$(29) \quad 0 \neq D_{\tau_k}^k(x, w) := \frac{\partial \dot{\phi}_{\tau_k}^k}{\partial q}.$$

If it turns out that $\tau_k \geq \pi_k$, we set $\tau_k := \pi_k$. Now distinguish the following cases (which should be checked sequentially):

1. $N_{\tau_k}^k := \{(x, w) \mid \phi_1^k(x, w) = \dots = \phi_{\tau_k}^k(x, w) = 0\} = \emptyset$. In this case we stop.
2. $\text{rank}(\partial \phi_{\tau_k}^k / \partial w) < \text{rank}((\partial \phi_{\tau_k}^k / \partial x)(\partial \phi_{\tau_k}^k / \partial w))$. In this case we stop.
3. $\pi_k = \tau_k$. In this case we stop, defining $N^* := N_{\pi_k}^k$.
4. If none of the cases 1, 2, or 3 hold, we define $\nu_{k+1} := \nu_k + \tau_k$ and $\hat{D}_{\tau_k}^k$ a vector of functions consisting of a maximal number of independent entries of $D_{\tau_k}^k$.

Then define the system

$$(30) \quad \Sigma_{ak+1} \begin{cases} \dot{x} = f(x) + g(x)u + p(x)q \\ \dot{w}_{i1} = w_{i2} \\ \dot{w}_{i2} = w_{i3} \\ \vdots \\ \dot{w}_{i\nu_{k+1}} = v_i \\ y_{k+1} = \begin{pmatrix} \phi_{\tau_k}^k \\ \hat{D}_{\tau_k}^k \end{pmatrix} \end{cases} \quad (i = 1, \dots, p).$$

Note that at every step k of Algorithm 3.8 we must apply only τ_k steps of the clamped dynamics algorithm to Σ_{ak-1} . If for a given x_0 at every step k there is a w_0^k

such that (x_0, w_0^k) is a regular point for the first τ_k steps of the clamped dynamics algorithm applied to Σ_{ak-1} , we call x_0 a *regular* point for Algorithm 3.8. If x_0 is a regular point, the algorithm terminates in a finite number, say k^* , of steps. Now we have the following theorem.

THEOREM 3.9. *Consider the system Σ . Let x_0 be a regular point for Algorithm 3.8 applied to Σ . Then*

1. *The (regular) DDDPdm is locally solvable around x_0 if and only if $k^* = 1$. Moreover, if $k^* = 1$, the algorithm terminates because of case 3.*

2. *If $k^* > 1$, the nDDDPdm is locally solvable around x_0 if and only if the algorithm terminates because of case 3.*

Proof. 1. See [17].

2. We only give the proof for the nDDDP. The proof for the nDDDPdm is analogous, with a few differences concerning the q -dependence of the compensator.

Sufficiency. Assume that the algorithm terminates because of case 3. For brevity of notation, we denote

$$(31) \quad \Phi(x, w) := \phi_{\pi_{k^*-1}}^{k^*-1}(x, w).$$

Thus, $N^* = \{(x, w) | \Phi(x, w) = 0\}$ and N^* is a controlled invariant submanifold for Σ_{ak^*-1} . Since case 2 does not apply for Φ , by the implicit function theorem we can find a partition (w_1, w_2) of w and a function $\Psi(x, w_2)$ such that locally N^* can be expressed as $\{(x, w) | w_1 = \Psi(x, w_2)\}$. Let $U \subset \mathcal{X}$ be a neighborhood of x_0 on which this parametrization of N^* holds. Now let $u = \alpha^*(x, w)$ be a control that renders N^* invariant for Σ_{ak^*-1} . Then it is obvious that if we apply this control to Σ_{ak^*-1} and we choose our initial conditions on N^* , we will have $y_i^{(j)}(t) = w_{ij}(t)$ for all t and thus $y(t), \dots, y^{(n)}(t)$ are independent of q for all t . Now let z_i ($i = 1, \dots, p$) be a vector of dimension ν_{k^*-1} and consider the compensator

$$(32) \quad Q \begin{cases} \dot{z}_i = Az_i + bv_i & (i = 1, \dots, p) \\ u = \alpha^*(x, z) \end{cases}$$

with (A, b) in Brunovsky form, initialized at $(\bar{x}, \Psi(\bar{x}, z^2), z^2)$ for an $\bar{x} \in U$. Then from the above it is clear that this compensator locally solves the nDDDP for Σ .

Necessity. Let x_0 be a regular point for Algorithm 3.8 applied to Σ . Assume that the nDDDP is locally solvable around x_0 by means of a compensator

$$(33) \quad Q \begin{cases} \dot{z} = \alpha(x, z) \\ u = \gamma(x, z). \end{cases}$$

Consider the system

$$(34) \quad \bar{\Sigma}_{a0} \begin{cases} \dot{x} = f(x) + g(x)u + p(x)q \\ \dot{w}_{i1} = w_{i2} \\ \dot{w}_{i2} = w_{i3} \\ \vdots \\ \dot{w}_{i\nu_{k^*-1}} = v_i \\ \bar{y}_{0i} = h_i(x) - w_{i1} \end{cases} \quad (i = 1, \dots, p).$$

Note that $\bar{\Sigma}_{a0}$ only differs from Σ_{a0} in the the number of w_{ij} 's for $\bar{\Sigma}_{a0}$ is larger. Since Q solves the nDDDP for Σ , it also solves the nDDDP for $\bar{\Sigma}_{a0}$. Thus, the clamped dynamics manifold of $\bar{\Sigma}_{a0} \circ Q$ is given by $M = \{(x, w, z) | \Phi(x, z) - w = 0\}$ for some

vector of functions $\Phi(x, z)$. Denote the clamped dynamics manifold of $\Sigma_{ak} \circ Q$ ($k = 0, \dots, k^* - 1$) by N^k . By Proposition 2.5, the functions determining the clamped dynamics manifold of Σ_{ak} are zero on N^k . This implies in particular that the functions determining the clamped dynamics manifold of Σ_{a0} are zero on M . Since $k^* > 1$, $\tau_0 < \pi_0$. Moreover, since $\partial \dot{\phi}_{\tau_0}^0 / \partial q \neq 0$ and

$$(35) \quad \dot{\phi}_{\tau_0}^0 = \frac{\partial \phi_{\tau_0}^0}{\partial x} [f(x) + g(x)u + p(x)q] + \frac{\partial \phi_{\tau_0}^0}{\partial w} \dot{w}$$

for $\bar{\Sigma}_{a0}$, it is easily seen that at least one of the functions determining M explicitly depends on q , and/or γ depends on q , unless $\hat{D}_{\tau_0}^0(x, w)$ and all of its time derivatives are zero on M . Since the former would be in contradiction with the form of M and/or the form of Q , the latter is the case. Now let the systems $\bar{\Sigma}_{ak}$ ($k = 1, \dots, k^* - 1$) be obtained from $\bar{\Sigma}_{ak-1}$ by augmenting the outputs according to

$$(36) \quad \bar{y}_k = \begin{pmatrix} \phi_{\tau_{k-1}}^{k-1} \\ \hat{D}_{\tau_{k-1}}^{k-1} \end{pmatrix}.$$

Again, note that $\bar{\Sigma}_{ak}$ only differs from Σ_{ak} in that the number of w_{ij} 's for $\bar{\Sigma}_{ak}$ is larger. Then, using the above and an induction argument, we can prove that the clamped dynamics manifold of $\bar{\Sigma}_{ak} \circ Q$ ($k = 0, \dots, k^* - 1$) is equal to M . If we compare the form of Σ_{ak} and $\bar{\Sigma}_{ak}$, we see that this implies that also N^k ($k = 0, \dots, k^* - 1$) is of the form $N^k = \{(x, w, z) \mid \Phi_k(x, z) - w = 0\}$ for some vector of functions $\Phi_k(x, z)$. Consider the sets $\tilde{N}^k := \{(x, w) \mid \text{there exists } z \text{ such that } (x, w, z) \in N^k\}$. By the form of N^k it is clear that $\tilde{N}^k \neq \emptyset$. Now, by Proposition 2.5, N^k can alternatively be written as $N^k = \{(x, w, z) \mid \phi_{\tau_k}^k(x, w) = 0, \psi_k(x, w, z) = 0\}$ for some vector of functions $\psi_k(x, w, z)$. Thus any element $(x, w) \in \tilde{N}^k$ must satisfy $\phi_{\tau_k}^k(x, w) = 0$, which implies that $\tilde{N}^k \subset N_{\tau_k}^k$. Hence $N_{\tau_k}^k \neq \emptyset$ and thus the algorithm cannot terminate because of case 1. Now assume that the algorithm terminates because of case 2. This implies that we cannot find a partition (w_1, w_2) of w and a vector of functions $\Psi_{k^*-1}(x, w_2)$ such that $N_{\tau_{k^*-1}}^{k^*-1} = \{(x, w) \mid w_1 = \Psi_{k^*-1}(x, w_2)\}$, i.e., x cannot be a "free variable" on $N_{\tau_{k^*-1}}^{k^*-1}$. By Proposition 2.5 this implies that x cannot be a free variable on N^{k^*-1} either, which gives a contradiction with the form of N^{k^*-1} . Hence the algorithm can only terminate because of case 3. \square

Remark 3.10. It is easy to see that while solving the nDDDPdm we can take the function $\alpha^*(x, q, w)$, which renders N^* as an invariant manifold for Σ_{ak^*-1} to be affine in q , i.e., we can take α^* of the form $\alpha^*(x, q, z) = \alpha_1^*(x, z) + \alpha_2^*(x, z)q$.

We chose the above form of the algorithm and construction of the compensator to make the algorithm and the proof of Theorem 3.9 as transparent as possible. However, the bookkeeping while applying the algorithm can become quite troublesome, and the order of the compensator can become unnecessarily high. For (relatively) simple examples, much can be improved by using ad hoc arguments in the vein of the algorithm, as will be illustrated by the following example.

Example 3.11. Consider the system

$$(37) \quad \begin{aligned} \dot{x}_1 &= x_2 u_1 + x_4, & y_1 &= x_1, \\ \dot{x}_2 &= x_3 + q_1, & y_2 &= x_3, \\ \dot{x}_3 &= x_1 u_1 + x_4, & y_3 &= x_5, \\ \dot{x}_4 &= x_5, & y_4 &= x_7, \end{aligned}$$

$$\begin{aligned}
\dot{x}_5 &= u_2 + x_9, \\
\dot{x}_6 &= x_8 u_3, \\
\dot{x}_7 &= x_6 u_2 + x_8 + x_9 + x_6 x_9, \\
\dot{x}_8 &= u_3, \\
\dot{x}_9 &= x_{10}, \\
\dot{x}_{10} &= q_2
\end{aligned}$$

((37) cont.)

for which we want to solve the nDDDP around points in the set $\{x \in \mathbb{R}^{10} \mid x_2 \neq 0, x_1 \neq 0, x_6 \neq 0, x_8 \neq 0\}$. Let us first restrict our attention to y_1 and y_2 . We find

$$\begin{aligned}
\dot{y}_1 &= x_2 u_1 + x_4 \Rightarrow u_1 = \frac{1}{x_2} (\dot{y}_1 - x_4), \\
\dot{y}_2 &= x_1 u_1 + x_4 = \frac{x_1}{x_2} (\dot{y}_1 - x_4) + x_4, \\
\ddot{y}_2 &= \frac{\dot{y}_1}{x_2} (\dot{y}_1 - x_4) - \frac{x_1(x_3 + q_1)}{x_2^2} (\dot{y}_1 - x_4) + \frac{x_1}{x_2} (\ddot{y}_1 - \dot{x}_5) + \dot{x}_5.
\end{aligned}$$

Thus \ddot{y}_2 intrinsically depends on q_1 . Hence we should have

$$(38) \quad \frac{x_1}{x_2^2} (\dot{y}_1 - x_4) = 0.$$

Since we are working in a neighborhood of point for which $x_1 \neq 0$, $x_2 \neq 0$, this implies that $(1/x_2)(\dot{y}_1 - x_4) = u_1 = 0$.

Having chosen $u_1 = 0$, we see from the structure of the system y_1 and y_2 can be made independent of q_2 if and only if y_3 and y_4 can be made independent of q_2 . Restricting our attention to y_3 and y_4 , we find

$$\begin{aligned}
\dot{y}_3 &= u_2 + x_0 \Rightarrow u_2 = \dot{y}_3 - x_9, \\
\dot{y}_4 &= x_6 u_2 + x_8 + x_9 + x_6 x_9 = x_6 \dot{y}_3 + x_8 + x_9, \\
\ddot{y}_4 &= x_8 \dot{y}_3 u_3 + x_6 \ddot{y}_3 + \dot{u}_3 + \dot{x}_{10} \Rightarrow u_3 = \frac{1}{x_8 \dot{y}_3 + 1} (\ddot{y}_4 - x_6 \ddot{y}_3 - \dot{x}_{10}).
\end{aligned}$$

(39)

Thus we can solve the nDDDPdm by choosing a compensator

$$\begin{aligned}
\dot{z} &= v_1, \\
u_1 &= 0, \\
u_2 &= z - x_9, \\
u_3 &= \frac{1}{x_8 z + 1} (v_2 - x_6 v_1 - x_{10}),
\end{aligned}$$

(40)

with $z(0) \neq -(1/x_8(0))$ and v_1, v_2 denoting the new inputs.

4. The nonlinear model matching problem. In this section we will use the results of the previous section to give a complete local solution of the nonlinear model matching problem. For linear systems, the idea that we can obtain a solution of the nonlinear model matching problem by associating it with a disturbance decoupling problem was elaborated in [32], [11]. It was first extended to nonlinear systems in [10].

First, we give a formulation of the problem. Consider a nonlinear plant P , described by equations of the form

$$(41) \quad P \begin{cases} \dot{x} = f(x) + g(x)u \\ y = h(x), \end{cases}$$

where $x \in \mathcal{X}$, an open subset of \mathbb{R}^n , the inputs $u \in \mathbb{R}^m$, the outputs $y \in \mathbb{R}^p$, f and h are vector-valued analytic functions, and g is a matrix-valued analytic function, all of appropriate dimensions.

Furthermore, let a nonlinear model M be given, which is described by the equations

$$(42) \quad M \begin{cases} \dot{x}^M = f^M(x^M) + g^M(x^M)u^M \\ y^M = h^M(x^M), \end{cases}$$

where $x^M \in \mathcal{X}^M$, an open subset of \mathbb{R}^{n_M} , the inputs $u^M \in \mathbb{R}^{m_M}$, the outputs $y^M \in \mathbb{R}^{p_M}$, f^M and h^M are vector-valued analytic functions, and g^M is a matrix-valued analytic function, all of appropriate dimensions.

The compensator Q used to control P is a nonlinear system described by equations of the form

$$(43) \quad Q \begin{cases} \dot{z} = a(x, z) + b(x, z)u^M \\ u = c(x, z) + d(x, z)u^M \end{cases}$$

with state $z \in \mathbb{R}^v$ and real analytic a, b, c, d .

The usual definition of the nonlinear model matching problem is given below (see [6], [10]).

DEFINITION 4.12 (Nonlinear model matching problem (MMP)). Given a plant $P = (f, g, h)$, a model $M = (f^M, g^M, h^M)$, and a point $(x_0, x_0^M) \in \mathcal{X} \times \mathcal{X}^M \subset \mathbb{R}^n \times \mathbb{R}^{n_M}$, find neighborhoods U of x_0 and U^M of x_0^M , an integer v , an open subset V of \mathbb{R}^v , a compensator $Q = (a, b, c, d)$ with a, b, c, d real analytic functions defined on $V \times U$, and a map $F: U \times U^M \mapsto V$, with the property that

$$(44) \quad y^{P \circ Q}(x, F(x, x^M), t) - y^M(x^M, t)$$

is independent of u^M for all t and for all $(x, x^M) \in U \times U^M$.

Given M and P we define an *extended system* E :

$$(45) \quad E \begin{cases} \dot{x}^E = f^E(x^E) + g^E(x^E)u^E + p^E(x^E)q^E \\ y^E = h^E(x^E) \end{cases}$$

with

$$x^E = \begin{pmatrix} x \\ x^M \end{pmatrix}, \quad f^E(x^E) = \begin{pmatrix} f(x) \\ f^M(x^M) \end{pmatrix}, \quad g^E(x^E) = \begin{pmatrix} g(x) \\ 0 \end{pmatrix},$$

$$p^E(x^E) = \begin{pmatrix} 0 \\ g^M(x^M) \end{pmatrix}, \quad h^E(x^E) = h(x) - h^M(x^M).$$

THEOREM 4.13. *The MMP is solvable for (M, P) if and only if the nDDDPdm is solvable for E .*

Proof. Necessity. Assume that the MMP is solvable for (M, P) by means of a compensator Q of the form (43). Then from the definition of the MMP it is obvious that the compensator

$$(46) \quad Q_E \begin{cases} \dot{z} = \alpha(x^E, q^E, z) \\ u = \gamma(x^E, q^E, z) \end{cases}$$

with $\alpha(x^E, q^E, z) = a(x, z) + b(x, z)q^E$, $\gamma(x^E, q^E, z) = c(x, z) + d(x, z)q^E$ solves the nDDDPdm for E .

Sufficiency (see also [10]). Assume that the nDDDPm is solvable for E by means of a compensator

$$(47) \quad Q_E \begin{cases} \dot{z} = \alpha_1(x^E, z) + \alpha_2(x^E, z)q^E \\ u = \gamma_1(x^E, z) + \gamma_2(x^E, z)q^E. \end{cases}$$

Note that we can indeed take the compensator to be affine in q (see Remark 3.10). Consider the following compensator for P :

$$(48) \quad Q \begin{cases} \dot{z}_1 = f^M(z_1) + g^M(z_1)u^M \\ \dot{z}_2 = \alpha_1(x, z_1, z_2) + \alpha_2(x, z_1, z_2)u^M \\ u = \gamma_1(x, z_1, z_2) + \gamma_2(x, z_1, z_2)u^M. \end{cases}$$

Then it is easy to see that Q solves the MMP for (M, P) . \square

Thus, using the results from Theorem 3.9, we can check if the MMP is solvable for given (M, P) and we can construct a compensator that solves the MMP. We will illustrate this by means of an example.

Example 4.14. Consider the plant

$$(49) \quad P \begin{cases} \dot{x}_1 = x_2 + x_2 u_2 & y_1 = x_1 \\ \dot{x}_2 = u_1 & y_2 = x_3 \\ \dot{x}_3 = u_2 & y_3 = x_3 \end{cases}$$

and the model

$$(50) \quad M \begin{cases} \dot{x}_1^M = x_2^M & y_1^M = x_1^M \\ \dot{x}_2^M = x_3^M + u_1^M & y_2^M = x_2^M \\ \dot{x}_3^M = u_2^M & y_3^M = x_4^M \\ \dot{x}_4^M = -x_4^M & \end{cases}$$

Then

$$\begin{aligned} \dot{y}_1 - \dot{y}_1^M &= x_2 + x_2 u_2 = x_2(1 + \dot{y}_3 - \dot{y}_3^M - x_4^M) - x_2^M, \\ \dot{y}_2 - \dot{y}_2^M &= u_1 - x_3^M - u_1^M \Rightarrow u_1 = \dot{y}_2 - \dot{y}_2^M + x_3^M + u_1^M, \\ \dot{y}_3 - \dot{y}_3^M &= u_2 + x_4^M \Rightarrow u_2 = \dot{y}_3 - \dot{y}_3^M - x_4^M, \\ \ddot{y}_1 - \ddot{y}_1^M &= (\dot{y}_2 - \dot{y}_2^M + x_3^M + u_1^M)(1 + \dot{y}_3 - \dot{y}_3^M - x_4^M) + x_2(\ddot{y}_3 - \ddot{y}_3^M - x_4^M) - x_3^M - u_1^M. \end{aligned}$$

Thus, $\ddot{y}_1 - \ddot{y}_1^M$ intrinsically depends on u_1^M (the disturbances). Hence to solve the MMP we must have

$$(51) \quad \dot{y}_3 - \dot{y}_3^M - x_4^M = u_2 = 0.$$

Then it is easy to see that we can solve the MMP by means of the nonregular compensator

$$(52) \quad Q \begin{cases} \dot{z}_1 = z_2 \\ \dot{z}_2 = z_3 + u_1^M \\ \dot{z}_3 = u_2^M \\ \dot{z}_4 = -z_4 \\ u_1 = z_3 + u_1^M \\ u_2 = 0 \end{cases}$$

with $z(0) = x^M(0)$.

5. Conclusions and remarks. In this paper we formulated and solved the local nDDDP. The solution was given by means of a constructive algorithm. A drawback

of the algorithm presented in this paper is that it does not give a compensator of minimal order solving the nDDDPm. Therefore we cannot make any statement about the solvability of the DDP by means of nonregular static state feedback. This remains a problem for future research. Another topic for future research is the generalization of the theory developed in this paper to general nonlinear systems of the form

$$(53) \quad \begin{cases} \dot{x} = f(x, q, u) \\ y = h(x, u). \end{cases}$$

Using the algorithm given in this paper we can immediately give sufficient conditions for the solvability of the nDDDPm for (53). Namely, consider the following extended system obtained from (53):

$$(54) \quad \begin{cases} \dot{x} = f(x, q, u) \\ \dot{u} = v \\ y = h(x, u). \end{cases}$$

Then any solution to the nDDDPm for (53) gives rise to a solution of the problem for (54). The converse is easily seen to be false. However, since for the system (53) versions of Singh's algorithm and the clamped dynamics algorithm are also available (see [26], [23], [24] for Singh's algorithm and [35] for the clamped dynamics algorithm), it seems that the method of this paper can be straightforwardly extended to systems of the form (53).

By proving that the solvability of the nonlinear MMP is equivalent to the solvability of an associated nDDDP with disturbance measurements, we also established a complete local solution of this problem. A problem that remains unsolved is the problem of internal stability of the compensated plant after we have solved the nonlinear MMP. Until now this problem has only been addressed in [3] in the case where the plant is a single-input, single-output system and in [15] in the case where the plant is decouplable by static state feedback. The problem consists in the fact that, even if we start from an internally stable plant and an internally stable model, we may very well introduce unstable unobservable modes in the closed loop. To solve this problem, further investigation of the structure of a model matching configuration is needed, especially concerning the "fixed" and "free" modes of such a configuration. For linear systems this investigation has already been performed in [31]. For nonlinear systems this is undoubtedly much more difficult to answer. So far, only results about "fixed" modes in the solution of the input-output decoupling problem have been obtained in [12]. It is not clear if a similar analysis is applicable for the nonlinear MMP considered here. We leave this open for future research.

Appendix.

Proof of Proposition 2.5. Denote the submanifolds obtained while applying the clamped dynamics algorithm to (6) and (6), (10) by $N_k = \{x \mid \phi_k(x) = 0\}$, M_k , respectively. We will prove by induction that we can find vectors of functions $\psi_k(x, z)$ such that $M_k = \{(x, z) \mid \xi_k(x, z) = 0\}$, with

$$\xi_k(x, z) = \begin{pmatrix} \phi_k(x) \\ \psi_k(x, z) \end{pmatrix}.$$

Obviously, we have $N_0 = \{x \mid \phi_0(x) = 0\}$, $M_0 = \{(x, z) \mid \xi_0(x, z) = 0\}$, with $\phi_0(x) = \xi_0(x, z) = h(x)$. Hence our claim holds for $k = 0$. Now apply the first step of the clamped

dynamics algorithm to (6), yielding matrices $A_1(x)$, $B_1(x)$, a vector of functions $\phi_1(x)$, and $N_1 = \{x \mid \phi_1(x) = 0\}$. Applying the first step of the clamped dynamics algorithm to (6), (10) yields

$$(55) \quad \begin{aligned} \dot{\xi}_0(x, z) &= \frac{\partial \phi_0}{\partial x}(x)[f(x) + g(x)(\gamma(x, z) + \delta(x, z)v)] \\ &= A_1(x) + B_1(x)\gamma(x, z) + B_1(x)\delta(x, z)v. \end{aligned}$$

Since (x_0, z_0) is a regular point for the clamped dynamics algorithm applied to (6), (10), $B_1(x)\delta(x, z)$ has full rank \tilde{r}_1 in a neighborhood of (x_0, z_0) in M_0 . It is clear that $\tilde{r}_1 < r_1 = \text{rank } B_1(x)$. Moreover, the rows of $\tilde{B}_1(x)\delta(x, z)$ are linearly dependent on the rows of $\tilde{B}_1(x)\delta(x, z)$, since the rows of $\tilde{B}_1(x)$ are linearly dependent on the rows of $\tilde{B}_1(x)$. Thus we can permute the entries of $\xi_0(x, z)$ in such a way that the first \tilde{r}_1 rows of $B_1(x)\delta(x, z)$ are linearly independent and the last $(p_0 - r_1)$ entries of $\xi_0(x, z)$ consist of $\dot{\phi}_0(x)$, i.e., we can write (55) as

$$(56) \quad \begin{pmatrix} \dot{\sigma}_0(x, z) \\ \dot{\sigma}_0(x, z) \\ \dot{\phi}_0(x, z) \end{pmatrix} = \begin{pmatrix} \tilde{A}_{\sigma_1}(x) + \tilde{B}_{\sigma_1}(x)\gamma(x, z) \\ \hat{A}_{\sigma_1}(x) + \hat{B}_{\sigma_1}(x)\gamma(x, z) \\ \hat{A}_1(x) + \hat{B}_1(x)\gamma(x, z) \end{pmatrix} + \begin{pmatrix} \tilde{B}_{\sigma_1}(x)\delta(x, z) \\ \hat{B}_{\sigma_1}(x)\delta(x, z) \\ \hat{B}_1(x)\delta(x, z) \end{pmatrix} v,$$

where $(\tilde{\sigma}_0^T, \hat{\sigma}_0^T)^T$ consists of the entries of $\tilde{\sigma}_0$ and $\tilde{B}_{\sigma_1}(x)\delta(x, z)$ has full row rank \tilde{r}_1 in a neighborhood of (x_0, z_0) in M_0 . Observe that

$$\begin{pmatrix} \tilde{B}_{\sigma_1}(x) \\ \hat{B}_{\sigma_1}(x) \end{pmatrix}$$

has full row rank r_1 in a neighborhood of (x_0, z_0) in M_0 . Thus, we can rewrite $\dot{\phi}_0(x, z)$ in (56) as

$$(57) \quad \dot{\phi}_0(x, z) = \hat{A}_1(x) + \hat{B}_1(x) \begin{pmatrix} \tilde{B}_{\sigma_1}(x) \\ \hat{B}_{\sigma_1}(x) \end{pmatrix}^+ \left[\begin{pmatrix} \dot{\sigma}_0(x, z) \\ \dot{\sigma}_0(x, z) \end{pmatrix} - \begin{pmatrix} \tilde{A}_{\sigma_1}(x) \\ \hat{A}_{\sigma_1}(x) \end{pmatrix} \right],$$

where

$$\begin{pmatrix} \tilde{B}_{\sigma_1} \\ \hat{B}_{\sigma_1} \end{pmatrix}^+ \text{ is a right inverse of } \begin{pmatrix} \tilde{B}_{\sigma_1} \\ \hat{B}_{\sigma_1} \end{pmatrix}.$$

Let $(\tilde{B}_{\sigma_1}(x)\delta(x, z))^+$ be a right inverse of $\tilde{B}_{\sigma_1}(x)\delta(x, z)$. Letting $v = -(\tilde{B}_{\sigma_1}(x)\delta(x, z))^+ \tilde{A}_{\sigma_1}(x)$ we find from (56)

$$(58) \quad \begin{pmatrix} \dot{\sigma}_0(x, z) \\ \dot{\sigma}_0(x, z) \end{pmatrix} = \begin{pmatrix} 0 \\ \hat{A}_{\sigma_1}(x) - \hat{B}_{\sigma_1}(x)(\tilde{B}_{\sigma_1}(x)\delta(x, z))^+ \tilde{A}_{\sigma_1}(x) \end{pmatrix} =: \begin{pmatrix} 0 \\ \bar{\sigma}_1(x, z) \end{pmatrix}.$$

Then M_1 is given as $M_1 = \{(x, z) \mid \phi_0(x) = 0, \dot{\phi}_0(x, z) = 0, \bar{\sigma}_1(x, z) = 0\}$. Since $\dot{\sigma}_0(x, z) = 0$ on M_1 and (obviously)

$$\begin{pmatrix} \tilde{B}_{\sigma_1}(x) \\ \hat{B}_{\sigma_1}(x) \end{pmatrix}^+ \begin{pmatrix} \tilde{A}_{\sigma_1}(x) \\ \hat{A}_{\sigma_1}(x) \end{pmatrix} = \tilde{B}_1^+(x) \tilde{A}_1(x),$$

we find from (57) that on M_1

$$(59) \quad \dot{\phi}_0(x, z) = \hat{A}_1(x) - \hat{B}_1(x) \tilde{B}_1^+(x) \tilde{A}_1(x) = \hat{\phi}_1(x).$$

Thus, $M_1 = \{(x, z) \mid \phi_0(x) = 0, \bar{\phi}_1(x) = 0, \bar{\sigma}_1(x, z) = 0\}$. Assume that $(\bar{\phi}_1^T(x), \bar{\sigma}_1^T(x, z))^T$ has constant rank \tilde{s}_1 in a neighborhood of (x_0, z_0) in M_1 . Obviously, $\tilde{s}_1 > s_1$. Since $\bar{\phi}_1(x)$ has constant rank s_1 in a neighborhood of (x_0, z_0) in M_1 , we can permute the entries of $\bar{\phi}_1(x)$ and $\bar{\sigma}_1(x, z)$ such that $(\bar{\phi}_{11}(x), \dots, \bar{\phi}_{1s_1}(x), \bar{\sigma}_{11}(x, z), \dots, \bar{\sigma}_{1\tilde{s}_1-s_1}(x, z))$

are independent on M_1 . Defining $\phi_1(x) = (\phi_0^T(x), \bar{\phi}_{11}(x), \dots, \bar{\phi}_{1s_1}(x))^T$, $\psi_1(x, z) = (\bar{\sigma}_{11}(x, z), \dots, \bar{\sigma}_{1s_1-s_1}(x, z))^T$ we find $M_1 = \{(x, z) | \phi_1(x) = 0, \psi_1(x, z) = 0\}$. Hence our claim also holds for $k = 1$. Using similar arguments as above, we can prove that our claim holds for $k = 0, \dots, k^*$, which completes the proof. \square

Acknowledgments. I would like to thank Henk Nijmeijer and Arjan van der Schaft for the many useful discussions we have had on the contents and organization of this paper.

REFERENCES

- [1] G. BASILE AND G. MARRO, *Controlled and conditioned invariant subspaces in linear system theory*, J. Optim. Theory Appl., 3 (1969), pp. 306–315.
- [2] S. P. BHATTACHARYYA, *Disturbance rejection in linear systems*, Internat. J. Control, 5 (1971), pp. 633–637.
- [3] C. I. BYRNES, R. CASTRO, AND A. ISIDORI, *Linear model matching with prescribed tracking error and internal stability for nonlinear systems*, in Analysis and Optimization of Systems, A. Bensoussan and J. L. Lions, eds., Lecture Notes in Control and Inform. Sci., 111, Springer-Verlag, Berlin, 1988, pp. 249–258.
- [4] C. I. BYRNES AND A. ISIDORI, *A frequency domain philosophy for nonlinear systems, with applications to stabilization and adaptive control*, in Proc. 23rd CDC, Las Vegas, NV, 1984, pp. 1569–1573.
- [5] ———, *Local stabilization of minimum-phase nonlinear systems*, Systems Control Lett., 11 (1988), pp. 9–17.
- [6] M. D. DI BENEDETTO, *A condition for the solvability of the nonlinear model matching problem*, in New Trends in Nonlinear Control Theory, J. Descusse, M. Fliess, A. Isidori, and D. Leborgne, eds., Lecture Notes in Control and Inform. Sci., 122, Springer-Verlag, Berlin, 1988, pp. 102–115.
- [7] ———, *New results on nonlinear model matching*, preprint, 1988.
- [8] M. D. DI BENEDETTO AND J. W. GRIZZLE, *An intrinsic notion of regularity for local output nulling, inversion and dynamic extension of nonsquare systems*, preprint, 1989.
- [9] M. D. DI BENEDETTO, J. W. GRIZZLE, AND C. H. MOOG, *Rank invariants of nonlinear systems*, SIAM J. Control Optim., 27 (1989) pp. 658–672.
- [10] M. D. DI BENEDETTO AND A. ISIDORI, *The matching of nonlinear models via dynamic state feedback*, SIAM J. Control Optim., 18 (1986), pp. 420–436.
- [11] E. EMRE AND M. L. J. HAUTUS, *A polynomial characterization of (A, B) -invariant and reachability subspaces*, SIAM J. Control Optim., 18 (1980), pp. 420–436.
- [12] A. ISIDORI AND J. W. GRIZZLE, *Fixed modes and nonlinear noninteracting control with stability*, IEEE Trans. Automat. Control, AC-33 (1988), pp. 907–914.
- [13] R. M. HIRSCHORN, *Invertibility of multivariable nonlinear control systems*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 855–865.
- [14] ———, *(A, B) -invariant distributions and disturbance decoupling of nonlinear systems*, SIAM J. Control Optim., 19 (1981), pp. 1–19.
- [15] H. J. C. HUIJBERTS, *Nonlinear model matching with an application to Hamiltonian systems*, IFAC Nonlinear Control Systems Design Symposium 1989, Capri, Italy, preprints.
- [16] ———, *Nonlinear model matching: A local solution and two worked examples*, in Proc. 1990 American Control Conference, San Diego, CA, 1990.
- [17] H. 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, this issue.
- [18] H. J. C. HUIJBERTS AND H. NIJMEIJER, *Local nonlinear model matching: From linearity to nonlinearity*, Automatica, 26 (1990), pp. 973–983.
- [19] A. ISIDORI, *The matching of a prescribed linear input-output behavior in a nonlinear system*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 258–265.
- [20] ———, *Nonlinear Control Systems*, 2nd ed., Springer-Verlag, Berlin, 1989.
- [21] A. ISIDORI, A. J. KRENER, C. GORI-GIORGI, AND S. MONACO, *Nonlinear decoupling via feedback: A differential geometric approach*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 331–345.
- [22] A. ISIDORI AND C. H. MOG, *On the nonlinear equivalent of the notion of transmission zeros*, in Modelling and Adaptive Control, C. I. Byrnes and A. Kurzhanski, eds., Lecture Notes in Control and Inform. Sci., 105, Springer-Verlag, Berlin, 1988, pp. 146–158.
- [23] U. KOTTA, *Right inverse of a discrete-time nonlinear system*, Internat. J. Control, 51 (1990), pp. 1–9.

- [24] U. KOTTA AND H. NIJMEIJER, *Dynamic disturbance decoupling for discrete time nonlinear systems* Proc. Academy USSR, Technical Cybernetics 1991, to appear. (In Russian.)
- [25] A. J. KRENER, *(Adf, g), (adf, g) and locally (adf, g) invariant and controllability distributions*, SIAM J. Control Optim., 23 (1985), pp. 523–549.
- [26] C.-W. LI AND Y.-K. FENG, *Functional reproducibility of general multivariable analytic nonlinear systems*, Internat. J. Control, 45 (1987), pp. 255–268.
- [27] M. MALABRE, *Structure à l'infini des triplets invariants. Application à la poursuite parfaite de modèle*, in Analysis and Optimization of Systems, A. Bensoussan and J. L. Lions, eds., Lecture Notes in Control and Inform. Sci., 44, Springer-Verlag, Berlin, 1982, pp. 43–53.
- [28] R. MARINO, *High-gain feedback in nonlinear control systems*, Internat. J. Control, 42 (1985), pp. 1369–1385.
- [29] C. H. MOOG, A. M. PERDON, AND G. CONTE, *Model matching and factorization for nonlinear systems: A structural approach*, SIAM J. Control Optim., 29 (1991), pp. 769–785.
- [30] B. C. MOORE AND L. M. SILVERMAN, *Model matching by state feedback and dynamic compensation*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 491–497.
- [31] A. S. MORSE, *Structure and design of linear model following systems*, IEEE Trans. Automat. Control, AC-18 (1973), pp. 346–354.
- [32] ———, *Minimal solutions to transfer matrix equations*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 131–133.
- [33] H. NIJMEIJER AND A. J. VAN DER SCHAFT, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.
- [34] S. N. SINGH, *A modified algorithm for invertibility in nonlinear systems*, IEEE Trans. Automat. Control, AC-26 (1990), pp. 595–598.
- [35] A. J. VAN DER SCHAFT, *On clamped dynamics of nonlinear systems*, in Analysis and Control of Nonlinear Systems, C. I. Byrnes, C. F. Martin, and R. E. Saeks, eds., Elsevier, Amsterdam, the Netherlands, 1988, pp. 499–506.
- [36] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, 3rd ed., Springer-Verlag, New York, 1985.

ON THE EXTENSION OF NEWTON'S METHOD TO SEMI-INFINITE MINIMAX PROBLEMS*

E. POLAK†, D. Q. MAYNE‡, AND J. E. HIGGINS†

Abstract. This paper introduces two new techniques for the analysis and construction of semi-infinite optimization algorithms. The first is a very simple technique for establishing the superlinear rate of convergence of semi-infinite optimization algorithms. The second technique enables specification of discretization rules that preserve the superlinear convergence of *conceptual* superlinearly converging semi-infinite optimization algorithms.

Natural extensions of Newton's method to semi-infinite optimization are used as a vehicle for presenting the techniques. In particular, it is shown that both local and global versions of the conceptual extension of Newton's method converge Q -superlinearly, with rate at least $3/2$, and that their implementations, based on the discretization rules, retain this rate of convergence.

Key words. Newton's method, minimax problems, nondifferentiable optimization, superlinear convergence

AMS(MOS) subject classifications. 49K35, 49M15, 49M39

1. Introduction. This is a dual-purpose paper. The first purpose of this paper is to introduce a novel technique for establishing the superlinear convergence of a class of semi-infinite optimization algorithms; the second is to demonstrate the degree to which various discretization effects, associated with semi-infinite optimization problems, can be taken into account. In particular, this paper introduces discretization rules that preserve the superlinear convergence of *conceptual* superlinearly converging semi-infinite optimization algorithms.

In his pioneering paper [29] dealing with perturbed Kuhn–Tucker points, Robinson showed that by applying the implicit function theorem to the first-order optimality conditions of a finitely constrained optimization problem and then relating the result to the search direction finding problem of a particular algorithm, we can sometimes establish the superlinear convergence of this algorithm. The same technique can also be used for establishing the superlinear convergence of finite minimax algorithms; see, e.g., [26].

Unfortunately, Robinson's technique cannot be used in conjunction with semi-infinite optimization algorithms because the assumptions of the implicit function theorem are not usually satisfied in the semi-infinite case. The technique in this paper is based on function approximations and is therefore not restricted by the linear independence requirements associated with techniques based on the implicit function theorem.

To illustrate both our new technique for establishing the superlinear convergence of a semi-infinite optimization algorithm and the manner in which discretization effects can be taken into account, we chose an extension of Newton's method for the solution of semi-infinite optimization problems. Our choice was motivated partly by the fact

* Received by the editors August 14, 1989; accepted for publication (in revised form) February 18, 1991. This research was sponsored in part by National Science Foundation grant ECS-8713334, Air Force Office of Scientific Research contract AFOSR-86-0116, and the State of California MICRO Program grant 532410-19900.

† Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory, University of California, Berkeley, California 94720.

‡ Department of Electrical Engineering, Imperial College of Science and Technology, London, SW7 2BT, Great Britain.

that Newton's method is the simplest method in the class that can be considered, and partly because Newton's method is one of the best understood, most studied, variously modified, adapted, and approximated algorithms in the literature (see, e.g., [6], [19], [9], [16], [17], [20], [25], [29], [30]).

In the area of nonlinear programming, Newton's method was at first used as a *local* method for unconstrained optimization of twice locally Lipschitz continuously differentiable, strongly convex functions on \mathbb{R}^n . Then, it was shown by Goldstein [8] that, for such functions, the local Newton method can be globally stabilized; i.e., it can be made globally convergent by the addition of the Armijo–Goldstein step-size rule [1], [8]. Since this step-size rule returns a stepsize of unity near a solution (see [8]), the Goldstein version of the globally stabilized Newton method converges quadratically. Finally, referring to [21], we see that it is possible to construct globally stabilized versions of Newton's method that converge quadratically in minimizing twice locally Lipschitz continuously differentiable, but not necessarily convex, functions on \mathbb{R}^n , whose local minimizers satisfy second-order sufficiency conditions. Such globally converging methods are obtained by using the Goldstein method if certain conditions are satisfied, and reverting to the Armijo gradient method [1] otherwise (see, e.g., [18]).

The extension of Newton's method (largely in the form of sequential quadratic programming) to semi-infinite optimization problems appears to have been confined to constrained problems that can be converted to ordinary nonlinear programming problems by means of the implicit function theorem (see, e.g., [11], [24], [5]). For example, a problem of the form

$$(1.1a) \quad \min \{f(x) \mid \phi(x, t) \leq 0, \forall t \in [0, 1]\}$$

can be converted to the standard nonlinear programming form

$$(1.1b) \quad \min \{f(x) \mid \phi(x, t^j(x)) \leq 0, j = 1, 2, \dots, q\},$$

when it is known that for all x near a local solution x^* , $\phi(x, \cdot)$ has exactly q local maximizers, and that $\phi_{tt}(x, t^j(x)) < 0$ for each j . It should be noted that some of these extensions are *conceptual algorithms* because in their analysis it was not considered that the local maximizers $t^j(x)$ cannot be computed exactly.

We can convert an unconstrained minimax problem of the form

$$(1.2a) \quad \min_{x \in \mathbb{R}^n} \max_{t \in [0, t]} \phi(x, t)$$

into a constrained problem of the form

$$(1.2b) \quad \min \{w \mid \phi(x, t) \leq w, \forall t \in [0, t]\},$$

and, assuming that the required assumptions are satisfied, apply one of the above-mentioned algorithms (i.e., [11], [24], [5]). Such an approach suffers from both aesthetic and practical drawbacks. First, it is displeasing to convert an unconstrained optimization problem into a constrained one. Second, to avoid the Maratos effect [15], we must use a curvilinear step-size rule or other modifications, such as the use of the modified Lagrangian merit function of Shittkowski and Powell, which are more complex than the simple Armijo–Goldstein rule mentioned earlier. Third, unlike Newton's method, the methods in [11], [24], [5] do not exhibit invariance under linear transformations. Last, we have observed that constrained semi-infinite optimization algorithms (such as Algorithm 5.7 in [22]) do not perform on (1.2b) as well as semi-infinite minimax algorithms (such as the version of Algorithm 5.2, based on (5.52) in [22]) do on (1.2a).

In this paper, we present natural extensions of both the local version of Newton's method and of the Goldstein globally stabilized version of Newton's method, for the solution of a class of convex semi-infinite minimax problems. The notable aspects of our work are (i) we do not impose the above-mentioned restrictive assumption that all the local maximizers of $\phi(x, \cdot)$ are strict and that their number is finite, (ii) we consider the most obvious approximations required to produce implementable algorithms, and (iii) we use a new and very simple technique for establishing superlinear convergence of our extensions of Newton's method. Since our technique is not based on the implicit function theorem (as in [29], [25]), it does not require the imposition of a strict complementarity condition.¹ In § 2 we show that a *conceptual* local Newton's method for semi-infinite minimax problems converges superlinearly with Q -rate $3/2$, under assumptions analogous to those needed in the minimization of twice locally Lipschitz continuously differentiable, strongly convex functions on \mathbb{R}^n . In § 3 we present a *conceptual* globally stabilized Newton method and show that it retains the Q -rate of $3/2$. In § 4 we present two *implementable* versions of the local Newton method for semi-infinite minimax problems and show that they converge locally with Q -rate $3/2$; a superlinearly converging (with Q -rate $3/2$) *implementable* version of our globally stabilized Newton method is presented in § 5. We present numerical results in § 6 and our concluding comments and final observations in § 7.

2. The local Newton method. We will consider the problem

$$(2.1a) \quad \mathbf{P}: \min_{x \in \mathbb{R}^n} \psi(x),$$

where

$$(2.1b) \quad \psi(x) = \max_{j \in \mathbf{q}} \max_{t \in [0,1]} \phi^j(x, t),$$

where $\mathbf{q} \triangleq \{1, 2, \dots, q\}$.

In keeping with standard assumptions for Newton's method (see [8]), we make the following hypotheses.

Assumption 2.1. (i) For all $j \in \mathbf{q}$, the functions $\phi^j: \mathbb{R}^n \times [0, 1]$ are twice Lipschitz continuously differentiable in the first argument (uniformly in the second).

(ii) For all $j \in \mathbf{q}$, $\phi^j(\cdot, \cdot)$, $\nabla_x \phi(\cdot, \cdot)$, and $\phi_{xx}^j(\cdot, \cdot)$ are all continuous.

(iii) There exist constants $0 < m \leq M$ such that for all $x \in \mathbb{R}^n$,

$$(2.2) \quad m \|h\|^2 \leq \langle h, \phi_{xx}^j(x, t)h \rangle \leq M \|h\|^2, \quad \forall h \in \mathbb{R}^n, \quad \forall t \in [0, 1], \quad \forall j \in \mathbf{q}.$$

Semi-infinite constrained optimization problems of the form $\min \{f(x) \mid \phi^j(x, t) \leq 0, \text{ for all } j \in \mathbf{q}, \text{ for all } t \in [0, 1]\}$ may be expressed in the form of \mathbf{P} by means of an appropriate exact penalty function. If $f(\cdot)$ is strictly convex and $\phi^j(\cdot, t)$ is convex for all $j \in \mathbf{q}$, and all $t \in [0, 1]$, then the resulting problem \mathbf{P} satisfies Assumption 2.1 (iii). Consequently, Assumption 2.1 is satisfied by many problems, particularly in the areas of optimal control and optimization-based design of linear multivariable control systems. For example, convex optimal control problems, such as the one described in § 6, when expressed in the form of \mathbf{P} satisfy Assumption 2.1 (iii). Also, some algorithms for the solution of more general state-constrained optimal control problems require an efficient subprocedure for solving convex optimal control problems that satisfy

¹ Referring to problem (1.2a), we note that an assumption of strict complementary slackness is highly restrictive for semi-infinite minimax problems because it implies that, at a solution x^* , the active gradients $\nabla_x \phi(x^*, t)$ are affinely independent which, in turn, implies that $\phi(x^*, \cdot)$ has at most $n+1$ maximizers. However, $\phi(x^*, \cdot)$ may well have a *continuum* of maximizers.

Assumption 2.1 (iii); see, for example, [28]. Another important class of optimization problems satisfying Assumption 2.1 (iii) arises in the design of linear multivariable control systems with control and state constraints when “ Q -parameterization” is employed; see [3].

PROPOSITION 2.1. *Suppose that Assumption 2.1 holds and that x^* is the minimizer of $\psi(\cdot)$. Then for all $x \in \mathbb{R}^n$,*

$$(2.3a) \quad \psi(x) - \psi(x^*) \geq \frac{m}{2} \|x - x^*\|^2.$$

Proof. For any $x \in \mathbb{R}^n$, and for any $j \in \mathbf{q}$, let

$$(2.3b) \quad \mathbf{q}^*(x) \triangleq \{j \in \mathbf{q} \mid \psi(x) = \max_{t \in [0,1]} \phi^j(x, t)\},$$

$$(2.3c) \quad T^{*j}(x) \triangleq \{t \in [0,1] \mid \phi^j(x, t) = \psi(x)\}.$$

Then, making use of the second-order expansion formula [7, p. 185] and of (2.2), we obtain that

$$\begin{aligned} \psi(x) - \psi(x^*) &\geq \max_{j \in \mathbf{q}} \max_{t \in [0,1]} \phi^j(x^*, t) - \psi(x^*) + \langle \nabla_x \phi^j(x^*, t), x - x^* \rangle + \frac{m}{2} \|x - x^*\|^2 \\ (2.3d) \quad &\geq \max_{j \in \mathbf{q}^*(x^*)} \max_{t \in T^{*j}(x^*)} \phi^j(x^*, t) - \psi(x^*) + \langle \nabla_x \phi^j(x^*, t), x - x^* \rangle + \frac{m}{2} \|x - x^*\|^2 \\ &= d\psi(x^*, x - x^*) + \frac{m}{2} \|x - x^*\|^2, \end{aligned}$$

where $d\psi(x^*, x - x^*)$ denotes the directional derivative of $\psi(\cdot)$ at x^* , in the direction $(x - x^*)$. Since x^* is the minimizer of $\psi(\cdot)$, $d\psi(x^*, x - x^*) \geq 0$, and (2.3a) follows. \square

By analogy with Newton's method for differentiable functions, we define a quadratic approximation $\hat{\psi}(\cdot | y)$ to $\psi(\cdot)$, around the point y , by

$$(2.4a) \quad \hat{\psi}(x | y) \triangleq \max_{j \in \mathbf{q}} \max_{t \in [0,1]} \phi^j(y, t) + \langle \nabla_x \phi^j(y, t), x - y \rangle + \frac{1}{2} \langle (x - y), \phi_{xx}^j(y, t)(x - y) \rangle.$$

Algorithm 2.1 (Local Newton Method).

Data: $x_0 \in \mathbb{R}^n$.

Step 0. Set $i = 0$.

Step 1. Compute

$$(2.4b) \quad x_{i+1} = \arg \min_{x \in \mathbb{R}^n} \hat{\psi}(x | x_i).$$

Step 2. Replace i by $i + 1$ and go to Step 1.

Proceeding as in the proof of Proposition 2.1, it can be shown that $\hat{\psi}(x | x_i) - \hat{\psi}(x_{i+1} | x_i) \geq \frac{1}{2} m \|x - x_{i+1}\|^2$. Hence we conclude that x_{i+1} is uniquely defined by (2.4b).

To establish the local convergence and rate of convergence of the above algorithm, we need the following lemmas.

LEMMA 2.1. *Suppose that Assumption 2.1 holds. Then there exists a $\hat{K} < \infty$ such that for any $x, y \in \mathbb{R}^n$,*

$$(2.5) \quad |\psi(x) - \hat{\psi}(x | y)| \leq \hat{K} \|x - y\|^3.$$

Proof. Let $L < \infty$ be a common Lipschitz constant for the Hessians $\phi_{xx}^j(\cdot, \cdot)$. Then, making use of second-order expansions, we obtain that

$$\begin{aligned}
 \psi(x) &= \max_{j \in q} \max_{t \in [0,1]} \phi^j(y, t) + \langle \nabla_x \phi^j(y, t)x, x - y \rangle + \frac{1}{2} \langle (x - y), \phi_{xx}^j(y, t)(x - y) \rangle \\
 (2.6) \quad &+ \int_{s \in [0,1]} (1 - s) \langle (x - y), [\phi_{xx}^j(y + s(x - y), t) - \phi_{xx}^j(y, t)](x - y) \rangle ds \\
 &\leq \hat{\psi}(x|y) + \frac{L}{6} \|x - y\|^3.
 \end{aligned}$$

The other half of the inequality in (2.5) follows similarly (with $\hat{K} = L/6$). \square

LEMMA 2.2. Suppose that Assumption 2.1 is satisfied. Let $h: \mathbb{R}^n \rightarrow \mathbb{R}^n$ and $\theta: \mathbb{R}^n \rightarrow \mathbb{R}$ be defined by

$$(2.7a) \quad h(x) \triangleq \arg \min_{h \in \mathbb{R}^n} \hat{\psi}(x + h|x),$$

$$(2.7b) \quad \theta(x) \triangleq \min_{h \in \mathbb{R}^n} \hat{\psi}(x + h|x) - \psi(x).$$

Then (a) both $h(\cdot)$ and $\theta(\cdot)$ are continuous; (b) for all $x \in \mathbb{R}^n$, $d\psi(x, h(x)) \leq \theta(x)$; (c) if x^* is a solution of (2.1a), then both $h(x^*) = 0$ and $\theta(x^*) = 0$; (d) for all $x \neq x^*$, $\theta(x) < 0$.

Proof. (a) Continuity of $\theta(\cdot)$ and $h(\cdot)$ follows from the maximum theorem in [2], strengthened by Assumption 2.1 (iii). The continuity of $h(\cdot)$ again follows from the maximum theorem in [2], which states that it is an upper-semicontinuous set-valued map, and the fact that $h(x)$ is always a singleton.

(b) Clearly, with $q^*(x)$, $T^{*j}(x)$ defined as in (2.3b), (2.3c), we must have that

$$(2.7c) \quad \theta(x) \geq \max_{j \in q^*(x)} \max_{t \in T^{*j}(x)} \phi^j(x, t) - \psi(x) + \langle \nabla_x \phi^j(x, t), h(x) \rangle = d\psi(x, h(x)).$$

Hence $d\psi(x, h(x)) \leq \theta(x)$.

(c) Since $0 \leq d\psi(x^*, h(x^*)) \leq \theta(x^*) \leq 0$, must hold, it follows that both $h(x^*) = 0$ and $\theta(x^*) = 0$.

(d) In view of Assumption 2.1 (iii), x^* is the only point satisfying the first-order necessary and sufficient condition $0 \in \partial\psi(x^*)$, where $\partial\psi(x)$ denotes the Clarke-generalized gradient of $\psi(\cdot)$ at x (for a definition, see [4, p. 27]). Hence this part is a simple generalization of Proposition 5.5 in [22], from which we see that $\theta(x) \leq 0$ for all $x \in \mathbb{R}^n$ and that $\theta(x) = 0$ if and only if $0 \in \partial\psi(x)$. \square

LEMMA 2.3. Suppose that $K \in (0, \infty)$, and that $t, s \geq 0$ are such that

$$(2.8a) \quad t^2 \leq K[(s + t)^3 + s^3],$$

$$(2.8b) \quad 0 \leq t \leq \frac{1}{9K}, \quad 0 \leq s \leq \frac{1}{9K}.$$

Then $t \leq s$ and

$$(2.8c) \quad t \leq 3\sqrt{K} s^{3/2}.$$

Furthermore, if $s \leq \gamma/9K$, with $\gamma \in (0, 1)$, then

$$(2.8d) \quad t \leq \sqrt{\gamma} s.$$

Proof. Let $\lambda \triangleq 1/9K$. Then, from (2.8a), (2.8b),

$$\begin{aligned}
 (2.9a) \quad t^2 &\leq K[2s^3 + 3s^2t + 3st^2 + t^3] \\
 &\leq K[2\lambda s^2 + 3\lambda s^2 + 3\lambda t^2 + \lambda t^2].
 \end{aligned}$$

Hence

$$(2.9b) \quad (1 - 4\lambda K)t^2 \leq 5K\lambda s^2.$$

Since $(1 - 4\lambda K) = 5K\lambda = 5/9$, it follows that $t \leq s$. Hence, replacing t by s in (2.8a), we obtain (2.8c).

Now, if $s \leq \gamma/9K$, then $\sqrt{s} \leq \sqrt{\gamma}/3\sqrt{K}$. Substituting for \sqrt{s} in (2.8c), we obtain (2.8d), which completes our proof. \square

COROLLARY 2.1. *Suppose that $K \in (0, \infty)$, $\gamma \in (0, 1)$, and that $\{\alpha_i\}_{i=0}^\infty$ is a real sequence such that*

$$(2.10a) \quad \alpha_{i+1}^2 \leq K[(\alpha_i + \alpha_{i+1})^3 + \alpha_i^3],$$

$$(2.10b) \quad 0 \leq \alpha_i \leq \frac{\gamma}{9K}, \quad \forall i \in \mathbb{N}.$$

Then $\alpha_i \rightarrow 0$ as $i \rightarrow \infty$ superlinearly, with Q -rate $3/2$.

Proof. It follows from Lemma 2.3 that $\alpha_{i+1} \leq \sqrt{\gamma} \alpha_i$, for all $i \in \mathbb{N}$. Hence $\alpha_i \rightarrow 0$ as $i \rightarrow \infty$. The $3/2$ Q -rate follows from (2.8c). \square

We are finally ready to establish the convergence properties of Algorithm 2.1.

THEOREM 2.1. *There exists a $\rho > 0$ such that if $\|x_0 - x^*\| \leq \rho$, where x^* is the solution of (2.1a), and $\{x_i\}_{i=0}^\infty$ is a sequence constructed by Algorithm 2.1, then, $x_i \rightarrow x^*$, as $i \rightarrow \infty$, Q -superlinearly, with rate at least $3/2$.*

Proof. Let $\alpha = m/2$, then, making use of (2.3a) and (2.4b), we obtain, for $i = 0, 1, 2, \dots$, that

$$(2.11a) \quad \begin{aligned} \alpha \|x_{i+1} - x^*\|^2 &\leq \psi(x_{i+1}) - \psi(x^*) \\ &\leq \psi(x_{i+1}) - \hat{\psi}(x_{i+1} | x_i) + \hat{\psi}(x_{i+1} | x_i) - \psi(x^*) \\ &\leq \psi(x_{i+1}) - \hat{\psi}(x_{i+1} | x_i) + \hat{\psi}(x^* | x_i) - \psi(x^*) \end{aligned}$$

because $\hat{\psi}(x_{i+1} | x_i) \leq \hat{\psi}(x^* | x_i)$, by construction of x_{i+1} . It now follows from (2.11a) and Lemma 2.1 that

$$(2.11b) \quad \begin{aligned} \|x_{i+1} - x^*\|^2 &\leq K[\|x_{i+1} - x_i\|^3 + \|x_i - x^*\|^3] \\ &\leq K[\|(x_{i+1} - x^*) - (x_i - x^*)\|^3 + \|x_i - x^*\|^3] \\ &\leq K[(\|x_{i+1} - x^*\| + \|x_i - x^*\|)^3 + \|x_i - x^*\|^3], \end{aligned}$$

where $K = \hat{K}/\alpha$ and \hat{K} is as in Lemma 2.1.

Next, since by Lemma 2.2, $h(\cdot)$ is continuous and $h(x^*) = 0$, it follows that given $\gamma^* \in (0, 1)$, there exists a $\bar{\rho} > 0$ such that if $\|x_i - x^*\| \leq \bar{\rho}$, then $\|h(x_i)\| = \|x_{i+1} - x_i\| \leq \gamma^*/18K$. Let $\rho^* = \min\{\bar{\rho}, \gamma^*/18K\}$. Then, if $\|x_i - x^*\| \leq \rho^*$, we must have that

$$(2.12) \quad \|x_{i+1} - x^*\| \leq \|x_{i+1} - x_i\| + \|x_i - x^*\| \leq \frac{\gamma^*}{18K} + \rho^* \leq \frac{\gamma^*}{9K}.$$

Letting $t \triangleq \|x_{i+1} - x^*\|$ and $s \triangleq \|x_i - x^*\|$, and making use of Lemma 2.3 (see (2.8b)), we obtain that

$$(2.13) \quad \|x_{i+1} - x^*\| \leq \sqrt{\gamma^*} \|x_i - x^*\|.$$

Hence, if $\|x_0 - x^*\| \leq \rho^*$, then $\|x_i - x^*\| \leq \rho^*$ for all $i = 1, 2, 3, \dots$, and therefore, by (2.13), $\|x_i - x^*\| \rightarrow 0$ as $i \rightarrow \infty$. It now follows from (2.11b) and Corollary 2.1 (via (2.8c)) that

$$(2.14) \quad \|x_{i+1} - x^*\| \leq 3\sqrt{K} \|x_i - x^*\|^{3/2}, \quad \forall i \in \mathbb{N},$$

which completes our proof. \square

3. The global Newton method. We will now present an extension of the globally stabilized Newton method, proposed by Goldstein in [7] (see also [23, p. 33]). Stabilization is achieved by adding an Armijo-type step-size rule to the local Newton method. The rate of convergence of the local Newton method is preserved because, as we will show, near the solution of (2.1a) under Assumption 2.1, the stepsize becomes unity; i.e., the global Newton method reverts to the local Newton method.

Algorithm 3.1 (Global Newton Method).

Data: $x_0 \in \mathbb{R}^n$, $\alpha, \beta \in (0, 1)$, $S \triangleq \{1, \beta, \beta^2, \dots\}$.

Step 0. Set $i = 0$.

Step 1. Compute $\theta(x_i)$, and $h_i = h(x_i)$, according to (2.7b), (2.7a). Stop if $\theta(x_i) = 0$.

Step 2. Compute the stepsize $\lambda_i \triangleq \max \{\lambda \in S \mid \psi(x_i + \lambda h_i) - \psi(x_i) \leq \lambda \alpha \theta(x_i)\}$.

Step 3. Set $x_{i+1} = x_i + \lambda_i h_i$. Replace i by $i + 1$ and go to Step 1.

First, we show that Algorithm 3.1 is globally convergent.

THEOREM 3.1. *Suppose that Assumption 2.1 holds and that x^* is the solution of (2.1a). Then any sequence $\{x_i\}_{i=0}^\infty$, constructed by Algorithm 3.1, converges to x^* .*

Proof. First, because of Assumption 2.1 (iii), the level sets of $\psi(\cdot)$ are bounded and, by construction in Step 2, $\psi(x_{i+1}) < \psi(x_i)$. Hence any sequence $\{x_i\}_{i=0}^\infty$, constructed by Algorithm 3.1, must have accumulation points. For the sake of contradiction, suppose that the sequence $\{x_i\}_{i=0}^\infty$ does not converge to x^* . Then it must have an accumulation point $x^{**} \neq x^*$. By Lemma 2.2, we then have that $\theta(x^{**}) < 0$ and $h(x^{**}) \neq 0$. Since by Lemma 2.2, the directional derivative $d\psi(x^{**}, h(x^{**})) \leq \theta(x^{**}) < 0$, it follows that there is an $s^{**} \in S$ such that

$$(3.1a) \quad \psi(x^{**} + s^{**}h(x^{**})) - \psi(x^{**}) < s^{**}\alpha\theta(x^{**}).$$

Hence, making use of the continuity of $\theta(\cdot)$ and $h(\cdot)$, for all x_i sufficiently near x^{**} , the stepsize $\lambda_i \geq s^{**}$ and $\theta(x_i) < \theta(x^{**})/2$. Therefore, for all such x_i ,

$$(3.1b) \quad \psi(x_i + \lambda_i h(x_i)) - \psi(x_i) \leq \lambda_i \alpha \theta(x_i) \leq s^{**} \alpha \theta(x^{**})/2.$$

Since the sequence $\{\psi(x_i)\}_{i=0}^\infty$ is monotone decreasing, (3.1b) implies that $\psi(x_i) \rightarrow -\infty$ as $i \rightarrow \infty$, which is a contradiction. Hence the theorem must be true. \square

Next, we establish superlinear convergence.

THEOREM 3.2. *Suppose that Assumption 2.1 holds and that x^* is the solution of problem (2.1a). Then any sequence $\{x_i\}_{i=0}^\infty$, constructed by Algorithm 3.1, converges to x^* , superlinearly, with Q -rate at least $3/2$.*

Proof. Since $\{x_i\}_{i=0}^\infty$ converges to x^* by Theorem 3.1, we only need to show that there exists an i_0 such that $\lambda_i = 1$ for all $i \geq i_0$, so that Algorithm 3.1 reduces to Algorithm 2.1, and invoke Theorem 2.1.

Now, it follows from (2.5) that

$$(3.2a) \quad \begin{aligned} \theta(x_i) &= \hat{\psi}(x_i + h(x_i) \mid x_i) - \psi(x_i + h(x_i)) + \psi(x_i + h(x_i)) - \psi(x_i) \\ &\geq \psi(x_i + h(x_i)) - \psi(x_i) - \hat{K} \|h(x_i)\|^3. \end{aligned}$$

Hence

$$(3.2b) \quad \psi(x_i + h(x_i)) - \psi(x_i) \leq \alpha \theta(x_i) + [(1 - \alpha)\theta(x_i) + \hat{K} \|h(x_i)\|^3].$$

Next, we establish a relationship between $\theta(x)$ and $\|h(x)\|$. Since $x + h(x)$ is the minimizer of $\hat{\psi}(\cdot \mid x)$, it follows that it satisfies the first-order condition

$$(3.3a) \quad 0 \in \partial \hat{\psi}(x + h(x) \mid x).$$

For any integer $p \geq 1$, let $\Sigma_p \triangleq \{\mu \in \mathbb{R}^p \mid \sum_{j=1}^p \mu^j = 1, \mu^j \geq 0, \text{ for all } j \in \mathbf{q}\}$. Then it follows from (3.3a), the definition of the generalized gradient $\partial \hat{\psi}(x + h(x) \mid x)$ (see [4]), and the Carathéodory theorem [31], that there exists a multiplier $\mu \in \Sigma_q$, multipliers $\nu_j \in \Sigma_{n+1}$, and $t_j^k \in [0, 1]$, with $j \in \mathbf{q}$ and $k \in \mathbf{n} + \mathbf{1}$, such that

$$(3.3b) \quad 0 = \sum_{j=1}^q \mu^j \sum_{k=1}^{n+1} \nu_j^k [\nabla_x \phi^j(x, t_j^k) + \phi_{xx}^j(x, t_j^k) h(x)],$$

which implies that²

$$(3.3c) \quad h(x) = - \left[\sum_{j=1}^q \mu^j \sum_{k=1}^{n+1} \nu_j^k \phi_{xx}^j(x, t_j^k) \right]^{-1} \sum_{j=1}^q \mu^j \sum_{k=1}^{n+1} \nu_j^k \nabla_x \phi^j(x, t_j^k).$$

Furthermore, the following complementary slackness condition (see (5.12a), (5.12b) in [22]) is satisfied:

$$(3.3d) \quad \begin{aligned} \theta(x) = & \sum_{j=1}^q \mu^j \sum_{k=1}^{n+1} \nu_j^k \left\{ [\phi^j(x, t_j^k) - \psi(x)] + \langle \nabla_x \phi^j(x, t_j^k), h(x) \rangle \right. \\ & \left. + \frac{1}{2} \langle h(x), \phi_{xx}^j(x, t_j^k) h(x) \rangle \right\}. \end{aligned}$$

Substituting for $h(x)$ from (3.3c) into (3.3d), we obtain, in view of Assumption 2.1 (iii), that

$$(3.4) \quad \begin{aligned} \theta(x) = & \sum_{j=1}^q \mu^j \sum_{k=1}^{n+1} \nu_j^k [\phi^j(x, t_j^k) - \psi(x)] \\ & - \frac{1}{2} \left\langle h(x), \left[\sum_{j=1}^q \mu^j \sum_{k=1}^{n+1} \nu_j^k \phi_{xx}^j(x, t_j^k) \right] h(x) \right\rangle \\ \leq & -\frac{m}{2} \|h(x)\|^2, \end{aligned}$$

with the last line following from the fact that $\phi^j(x, t_j^k) - \psi(x) \leq 0$ for all t_j^k .

Substituting for $\theta(x)$ from (3.4) into (3.2b), we obtain

$$(3.5a) \quad \psi(x_i + h(x_i)) - \psi(x_i) \leq \alpha \theta(x_i) - [m(1 - \alpha)/2 - \hat{K} \|h(x_i)\|] \|h(x_i)\|^2.$$

Since $h(x_i) \rightarrow 0$ as $i \rightarrow \infty$, it follows that there exists an i_0 such that for all $i \geq i_0$,

$$(3.5b) \quad \psi(x_i + h(x_i)) - \psi(x_i) \leq \alpha \theta(x_i),$$

i.e., that $\lambda_i = 1$. This completes our proof. \square

4. Implementations of the local algorithm. Note that numerical evaluations of $\psi(x)$ and $\theta(x)$, and hence of $h(x)$, are only approximate for $\psi(x)$ because intervals must be discretized, and for $\theta(x)$ and $h(x)$, because they are defined by a convex optimization problem that can only be solved approximately. Hence both the local and the global Newton methods that we have presented (Algorithms 2.1 and 3.1, respectively) must be viewed as *conceptual*. This brings us to the question of whether it is possible to construct *implementable* algorithms, using some form of discretization of the interval

² Since the $\mu^j \geq 0$ and the $\nu_j^k \geq 0$ in (3.3c), it follows from (2.2) that the matrix

$$\left[\sum_{j=1}^q \mu^j \sum_{k=1}^{n+1} \nu_j^k \phi_{xx}^j(x, t_j^k) \right]$$

is invertible.

$[0, 1]$, appearing in (2.1b), as well as some truncation rule for the algorithm used in computing approximations to $\theta(x)$, which retain the basic properties of Algorithms 2.1 and 3.1.

We need to strengthen Assumption 2.1 by adding the following hypothesis.

Assumption 4.1. There exists a Lipschitz constant $L < \infty$,³ such that for all $x \in \mathbb{R}^n$,

$$(4.1a) \quad |\phi^j(x, t) - \phi^j(x, t')| \leq L|t - t'|, \quad \forall t, t' \in [0, 1], \quad \forall j \in \mathbf{q}.$$

$$(4.1b) \quad \|\nabla_x \phi^j(x, t) - \nabla_x \phi^j(x, t')\| \leq L|t - t'|, \quad \forall t, t' \in [0, 1], \quad \forall j \in \mathbf{q}.$$

$$(4.1c) \quad \|\phi_{xx}^j(x, t) - \phi_{xx}^j(x, t')\| \leq L|t - t'|, \quad \forall t, t' \in [0, 1], \quad \forall j \in \mathbf{q}.$$

We begin with the following observations. For any integer $N > 0$, let⁴

$$(4.2a) \quad T_N \triangleq \{t \mid t = \frac{k}{N}, k = 0, 1, 2, \dots, N\},$$

$$(4.2b) \quad \psi_N(x) \triangleq \max_{j \in \mathbf{q}} \max_{t \in T_N} \phi^j(x, t),$$

$$(4.2c) \quad \hat{\psi}_N(x|y) \triangleq \max_{j \in \mathbf{q}} \max_{t \in T_N} \phi^j(y, t) + \langle \nabla_x \phi^j(y, t), x - y \rangle + \frac{1}{2} \langle (x - y), \phi_{xx}^j(y)(x - y) \rangle,$$

$$(4.2d) \quad h_N(x) \triangleq \arg \min_{h \in \mathbb{R}^n} \hat{\psi}_N(x + h|x),$$

$$(4.2e) \quad \theta_N(x) \triangleq \min_{h \in \mathbb{R}^n} \hat{\psi}_N(x + h|x) - \psi_N(x).$$

The relationships between the quantities associated with the original problem P in (2.1a) and the approximating problems

$$(4.3) \quad P_N : \min_{x \in \mathbb{R}^n} \psi_N(x),$$

are shown in the following result.

PROPOSITION 4.1. Suppose that Assumptions 2.1 and 4.1 hold. Let x^* denote the solution of (2.1), and, for any positive integer N , let x_N^* denote the solution of the discretized problem P_N . Then

$$(4.4a) \quad |\psi(x^*) - \psi_N(x_N^*)| \leq \frac{L}{2N},$$

$$(4.4b) \quad \|x^* - x_N^*\|^2 \leq \frac{2L}{mN},$$

and, for every bounded set $B \subset \mathbb{R}^n$, there exists an $L' < \infty$ such that

$$(4.4c) \quad |\theta(x) - \theta_N(x)| \leq \frac{L'}{2N}, \quad \forall x \in B,$$

$$(4.4d) \quad \|h(x) - h_N(x)\|^2 \leq \frac{L'}{mN}, \quad \forall x \in B.$$

Proof. First, let $x \in \mathbb{R}^n$ be arbitrary. Then, because $T_N \subset [0, 1]$,

$$(4.5a) \quad \frac{-L}{2N} < 0 \leq \psi(x) - \psi_N(x).$$

³ At the expense of some complication, it is possible to carry out the following analysis using local, rather than global, Lipschitz constants.

⁴ Note that there is nothing special about the discretization (4.2a). Any family of discrete sets $T_N \subset T$, where $T \triangleq [0, 1]$ can be used provided that (i) $d(T_N, T) \rightarrow 0$ as $N \rightarrow \infty$, and (ii) for any sequence of integers $\{N_i\}_{i=0}^\infty$, such that $N_{i+1} \geq 2N_i$, $\sum d(T_{N_i}, T) < \infty$, where $d(\cdot, \cdot)$ denotes the Hausdorff distance.

Next, let $\hat{t} \in [0, 1]$ and $\hat{j} \in \mathbf{q}$ be such that $\psi(x) = \phi^{\hat{j}}(x, \hat{t})$. Then there exist points $t' \in T_N$, such that $|t' - \hat{t}| \leq 1/(2N)$ and hence, making use of (4.1),

$$(4.5b) \quad \psi_N(x) \geq \phi^{\hat{j}}(x, t') \geq \phi^{\hat{j}}(x, \hat{t}) - \frac{L}{2N} = \psi(x) - \frac{L}{2N}.$$

Thus we have shown that

$$(4.5c) \quad |\psi(x) - \psi_N(x)| \leq \frac{L}{2N}.$$

As a result, we have

$$(4.5d) \quad \psi(x^*) \leq \psi(x_N^*) \leq \psi_N(x_N^*) + \frac{L}{2N}$$

and

$$(4.5e) \quad \psi_N(x_N^*) \leq \psi_N(x^*) \leq \psi(x^*),$$

which gives us (4.4a).

Next, making use of (2.3a) and (4.5c), we obtain that

$$(4.6a) \quad \frac{m}{2} \|x_N^* - x^*\|^2 \leq \psi(x_N^*) - \psi(x^*) \leq \psi_N(x_N^*) + \frac{L}{2N} - \psi(x^*) \leq \frac{L}{2N} < \frac{L}{N},$$

which establishes (4.4b).

Now suppose $B \subset \mathbb{R}^n$ is bounded, and let $x \in B$. Let the functions $\eta_x^j: \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}$, $j \in \mathbf{q}$, be defined by

$$(4.6b) \quad \eta_x^j(h, t) \triangleq \phi^j(x, t) + \langle \nabla_x \phi^j(x, t), h \rangle + \frac{1}{2} \langle h, \phi_{xx}^j(x, t) h \rangle.$$

Let $G \triangleq \max_{x \in B} \max_{j \in \mathbf{q}} \max_{t \in [0, 1]} \|\nabla_x \phi^j(x, t)\|$. Then it follows from the inequality

$$(4.6c) \quad \begin{aligned} \max_{j \in \mathbf{q}} \max_{t \in [0, 1]} \eta_x^j(h, t) &\geq \max_{j \in \mathbf{q}} \max_{t \in [0, 1]} \phi^j(x, t) - G \|h\| + \frac{m}{2} \|h\|^2 \\ &= \psi(x) - G \|h\| + \frac{m}{2} \|h\|^2 \end{aligned}$$

that if $\|h\| > 2G/m$, then the left-hand side of (4.6c) is greater than $\psi(x)$. Since $\theta(x) \leq 0$, we must have $\|h(x)\| \leq 2G/m$. A similar analysis shows that $\|h_N(x)\| \leq 2G/m$, also. Now suppose that $\|h\| \leq 2G/m$, and let $L_B \triangleq L(1 + 2G/m + 2G^2/m^2)$. Then

$$(4.6d) \quad |\eta_x^j(h, t) - \eta_x^j(h, t')| \leq L_B |t - t'|, \quad \forall t, t' \in [0, 1], \quad \forall j \in \mathbf{q}.$$

It now follows from (2.7b), (4.2e), and (4.5c), and an argument similar to that used to establish (4.4a), (4.4b) that (4.4c), (4.4d) hold, with $L' \triangleq L_B + L$. \square

Comment 4.1. Note that it follows from duality theory [14] that because (4.2e) is a convex problem, the dual of (4.2e) is given by

$$(4.7a) \quad \theta_N(x) = \max_{\mu \in \Sigma_N} J_d(\mu), \quad \Sigma_N \triangleq \left\{ \mu \in \mathbb{R}^q \times \mathbb{R}^{N+1} \left| \sum_{\substack{j \in \mathbf{q} \\ t \in T_N}} \mu^{j,t} = 1, \mu^{j,t} \geq 0 \right. \right\},$$

where

$$(4.7b) \quad J_{d,N}(\mu) \triangleq - \sum_{\substack{j \in \mathbf{q} \\ t \in T_N}} \mu^{j,t} [\psi_N(x) - \phi^j(x, t)] \\ - \frac{1}{2} \left\langle \sum_{\substack{j \in \mathbf{q} \\ t \in T_N}} \mu^{j,t} \nabla_x \phi^j(x, t), \left(\sum_{\substack{j \in \mathbf{q} \\ t \in T_N}} \mu^{j,t} \phi_{xx}^j(x, t) \right)^{-1} \sum_{\substack{j \in \mathbf{q} \\ t \in T_N}} \mu^{j,t} \nabla_x \phi^j(x, t) \right\rangle.$$

Hence, given any set of admissible multipliers $\mu^{j,t}$, we have that $\theta_N(x) \geq J_{d,N}(\mu)$.

Now an algorithm such as the barrier function method in [27], applied to (4.2e), produces not only approximations \hat{h} to $h_N(x)$, but also associated multipliers $\mu^{j,t}$, while an algorithm such as the Levitin–Polyak method [13], applied to (4.7a) produces multipliers $\mu^{j,t}$ that, via (3.3c) can be used to obtain an approximation \hat{h} to $h_N(x)$. Either way, we have that

$$(4.7c) \quad J_{p,N}(\hat{h}) \triangleq \hat{\psi}_N(x + \hat{h}) - \psi_N(x) \geq \theta_N(x) \geq J_{d,N}(\mu).$$

Therefore, given any $\varepsilon > 0$, to determine when such an algorithm has constructed an approximation $h_{N,\varepsilon}(x)$ such that

$$(4.7d) \quad 0 \leq \hat{\psi}_N(x + h_{N,\varepsilon}(x) | x) - \psi_N(x) - \theta_N(x) \leq \varepsilon,$$

we only need to check whether $J_{p,N}(h_{N,\varepsilon}(x)) - J_{d,N}(\mu) \leq \varepsilon$. Hence we see that the construction of such $h_{N,\varepsilon}(x)$ is a finite process. Furthermore, it follows from Proposition 2.1, applied to the function $h \mapsto \hat{\psi}_N(x + h | x)$ and (4.7d), that

$$(4.7e) \quad \varepsilon \geq \hat{\psi}_N(x + h_{N,\varepsilon}(x) | x) - \hat{\psi}_N(x + h_N(x) | x) \geq \frac{m}{2} \|h_{N,\varepsilon}(x) - h_N(x)\|^2.$$

We can now follow one of two alternatives. The first is to decide on an acceptable level of error and then to use (4.4a) or (4.4b) to determine the required level of discretization, i.e., the parameter N . In that case, we propose to solve P_N and we only need to invent a scheme for truncating the computation of $h_N(x)$. Such a scheme is incorporated in the following implementation of the local Newton method for solving problems P_N . The second alternative involves increasing the discretization mesh progressively, rather than using a fixed discretization.. This second alternative will be discussed subsequently.

Algorithm 4.1 (Implementable Finite Minimax Local Newton Method for P_N).

Data: $x_0 \in \mathbb{R}^n$, $\hat{\varepsilon} \in (0, 1)$.

Step 0: Set $i = 0$.

Step 1. Set $\varepsilon = \hat{\varepsilon}$.

Step 2. Compute a vector $h_{N,\varepsilon}(x_i) \in \mathbb{R}^n$ such that⁵

$$(4.8a) \quad 0 \leq \hat{\psi}_N(x_i + h_{N,\varepsilon}(x_i) | x_i) - \psi_N(x_i) - \theta_N(x_i) \leq \varepsilon.$$

Step 3. If

$$(4.8b) \quad \hat{\psi}_N(x_i + h_{N,\varepsilon}(x_i) | x_i) - \psi_N(x_i) \leq -2\varepsilon$$

and

$$(4.8c) \quad \varepsilon \leq \|h_{N,\varepsilon}(x_i)\|^3,$$

set $x_{i+1} = x_i + h_{N,\varepsilon}(x_i)$, $\varepsilon_i = \varepsilon$,⁶ and go to Step 4. Else replace ε by $\varepsilon/2$ and go to Step 2.

Step 4. Replace i by $i + 1$ and go to Step 1.

⁵ Note that $\theta_N(x_i)$ is *not* evaluated. See the paragraph preceding (4.7d).

⁶ Note that the computation of ε_i need not always begin with $\hat{\varepsilon}$. Rather, it is more efficient to start with ε_{i-1} .

Comment 4.2. The structure of the tests (4.8a)–(4.8c) is dictated by the proofs to follow, which establish the $3/2$ rate of convergence of the algorithm. Note that (4.8b) ensures that

$$(4.8d) \quad \varepsilon_i \leq \frac{-\theta_N(x_i)}{2}.$$

Hence, since $\theta_N(x_i) \rightarrow 0$ as $x_i \rightarrow x_N^*$, it follows that Algorithm 4.1 computes approximates $h_N(x_i)$ more accurately as the solution of P_N is approached. Also, if Algorithm 4.1 is initialized with $x_0 = x^*$, it cycles indefinitely up in the loop defined by Steps 2 and 3, reducing ε to zero.

THEOREM 4.1. *There exists a $\rho > 0$ such that if $\|x_0 - x_N^*\| \leq \rho$, where x_N^* is the solution of (4.3), and $\{x_i\}_{i=0}^\infty$ is a sequence constructed by Algorithm 4.1, then, $x_i \rightarrow x_N^*$, as $i \rightarrow \infty$, Q -superlinearly, with rate at least $3/2$.*

Proof. First, we note that (2.5) holds with $\psi(\cdot)$, $\hat{\psi}(\cdot|\cdot)$ replaced by $\psi_N(\cdot)$, $\hat{\psi}_N(\cdot|\cdot)$, respectively, that we may assume that $K \geq 1$ in (2.5), and that Theorem 2.1 equally applies to the obvious simplification of the local Newton method for problem P_N .

Next, for any i , let $x'_{i+1} \triangleq x_i + h_N(x_i)$. Then, by (2.11b),

$$(4.9a) \quad \|x'_{i+1} - x_N^*\|^2 \leq K[\|x'_{i+1} - x_i\|^3 + \|x_i - x_N^*\|^3].$$

Since by (4.7e) and (4.8c), we have that

$$(4.9b) \quad \begin{aligned} \|x_{i+1} - x'_{i+1}\|^2 &= \|h_{N,\varepsilon_i}(x_i) - h_N(x_i)\|^2 \\ &\leq \frac{2\varepsilon_i}{m} \leq \frac{2}{m} \|h_{N,\varepsilon_i}(x_i)\|^3 = \frac{2}{m} \|x_{i+1} - x_i\|^3, \end{aligned}$$

we obtain, using (4.9a) and the fact that $\|x + y\|^p \leq 2^{p-1}[\|x\|^p + \|y\|^p]$, $p = 2, 3$, that, with $K \geq 1$,

$$(4.9c) \quad \begin{aligned} \|x_{i+1} - x_N^*\|^2 &\leq 2[\|x'_{i+1} - x_N^*\|^2 + \|x_{i+1} - x'_{i+1}\|^2] \\ &\leq 2K[\|x'_{i+1} - x_i\|^3 + \|x_i - x_N^*\|^3 + \|x_{i+1} - x'_{i+1}\|^2] \\ &\leq 8K[\|x_{i+1} - x_i\|^3 + \|x_{i+1} - x'_{i+1}\|^3 + \|x_i - x_N^*\|^3 + \|x_{i+1} - x'_{i+1}\|^2]. \end{aligned}$$

Assuming that $x_i - x_N^*$ is sufficiently small, we must have, in view of the fact that by Lemma 2.1 $\theta_N(x_i) \rightarrow 0$ as $x_i \rightarrow x_N^*$ and, by (4.8d), that $2\varepsilon_i/m < 1$ and hence, by (4.9b), that $\|x_{i+1} - x'_{i+1}\| < 1$. Therefore, making use of (4.9b), (4.9c) leads to the conclusion that there exists a $K' \in [16K, \infty)$, depending on m , such that (4.9c) reduces to

$$(4.10a) \quad \begin{aligned} \|x_{i+1} - x_N^*\|^2 &\leq 16K[\|x_{i+1} - x_i\|^3 + \|x_i - x_N^*\|^3 + \|x_{i+1} - x'_{i+1}\|^2] \\ &\leq K'[\|x_{i+1} - x_i\|^3 + \|x_i - x_N^*\|^3]. \end{aligned}$$

The proof can now be completed by using arguments similar to those following (2.11b) in the proof of Theorem 2.1. This requires that we show that given any $\delta > 0$, there exists a $\rho > 0$ such that if $\|x_i - x_N^*\| \leq \rho$, then $\|x_{i+1} - x_i\| \leq \delta$. Making use of the triangle inequality, (4.7d) and (4.8d), we obtain that

$$(4.10b) \quad \begin{aligned} \|x_{i+1} - x_i\| &= \|h_{N,\varepsilon_i}(x_i)\| \\ &\leq \|h_{N,\varepsilon_i}(x_i) - h_N(x_i)\| + \|h_N(x_i)\| \\ &\leq \sqrt{\frac{-2\theta_N(x_i)}{m}} + \|h_N(x_i)\|. \end{aligned}$$

The desired continuity result now follows from Lemma 2.2, and we can now proceed as in the proof of Theorem 2.1, following (2.11b), to complete this proof. \square

There is evidence in the literature (see, e.g., [12], [10]) that we can reduce computing times considerably by increasing the discretization meshsize progressively, rather than using the finest mesh from the very start. This idea is incorporated in the following implementation of the local Newton method, which adjusts both the precision with which successive iterates are computed, as well as the meshsize.

Algorithm 4.2 (Implementable Finite Minimax Local Newton Method for P).

Data. $x_0 \in \mathbb{R}^n$, $\hat{\varepsilon} \in (0, 1)$, $K, \hat{\tau} \ll 1$, $N_0 \in \mathbb{N}$.

Step 0. Set $i = 0$.

Step 1. Set $\varepsilon = \hat{\varepsilon}$, $\tau = \hat{\tau}$, $N = N_i$.⁷

Step 2. Compute a vector $h_{N,\varepsilon}(x_i) \in \mathbb{R}^n$ such that (see Comment 4.1)

$$(4.11a) \quad 0 \leq \hat{\psi}_N(x_i + h_{N,\varepsilon}(x_i) | x_i) - \psi_N(x_i) - \theta_N(x_i) \leq \varepsilon.$$

Step 3. If

$$(4.11b) \quad \hat{\psi}_N(x_i + h_{N,\varepsilon}(x_i) | x_i) - \psi_N(x_i) \leq -2\varepsilon$$

and

$$(4.11c) \quad \varepsilon \leq \|h_{N,\varepsilon}(x_i)\|^3,$$

set $x'_{i+1} = x_i + h_{N,\varepsilon}(x_i)$ and go to Step 4.

Else,

if $\varepsilon \geq \tau$, replace ε by $\varepsilon/2$ and go to Step 2,

else, replace τ by $\tau/2$, N by $2N$, and go to Step 2.

Step 4. If

$$(4.11d) \quad \frac{K}{N} \leq \|x'_{i+1} - x_i\|^3,$$

set $x_{i+1} = x_{i+1}$, $\varepsilon_i = \varepsilon$, $N_i = N$, and go to Step 5. Else, replace N by $2N$ and go to Step 2.

Step 5. Replace i by $i+1$ and go to Step 2.

Comment 4.3. (a) The function of the coefficient K in (4.11d) is to limit the growth of the discretization parameter N . Thus suppose that we are willing to accept a solution corresponding to N^* discretization points, and that our stopping criterion is $\|h_{N^*}(x_i)\| \leq \omega$, with $\omega \ll 1$. Then we would set $K \leq N^* \omega$.

(b) It is possible that Algorithm 4.2 may cycle indefinitely in the loop defined by Step 2–Step 3 or Step 2–Step 4. If this happens, however, then $x_i = x^*$. To see this, note that if this cycling occurs, then one of the tests (4.11b)–(4.11d) must fail infinitely often and as a consequence, $N \rightarrow \infty$ and $\varepsilon \rightarrow 0$. Combining (4.4d) and (4.7e), we have $h_{N,\varepsilon}(x_i) \rightarrow h(x_i)$. Now suppose that the test (4.11b) fails infinitely often. The inequalities (4.4c), (4.5c), and (4.7e) imply that $\hat{\psi}(x_i + h(x_i) | x_i) - \psi(x_i) = 0$, and it follows from Lemma 2.2 that $x_i = x^*$. If either of the other two inequalities (4.11c)–(4.11d) fails infinitely often, it follows immediately that $h_{N,\varepsilon}(x_i) \rightarrow 0$, and so $h(x_i) = 0$, from which the desired conclusion follows.

THEOREM 4.2. *There exists a $\rho > 0$ and an integer $N_0 < \infty$, such that if $\|x_0 - x^*\| \leq \rho$, where x^* is the solution of (2.1a), and $\{x_i\}_{i=0}^\infty$ is a sequence constructed by Algorithm 4.2, then $x_i \rightarrow x^*$ as $i \rightarrow \infty$, Q -superlinearly, with rate at least $3/2$.*

⁷ Although it is reasonable to key ε_i , which controls the precision with which $\theta_{N_i}(x_i)$ is approximated, to the actual value of $\theta_{N_i}(x_i)$, so that ε_i may or may not decrease monotonically, it makes better sense to increase the discretization parameter N_i monotonically.

Proof. First, assuming that $\|x_i - x_{N_i}^*\| \leq 1$, it follows from (4.10) that for some $K' < \infty$, independent of N_i ,

$$(4.12a) \quad \|x_{i+1} - x_{N_i}^*\|^2 \leq K'[\|x_{i+1} - x_i\|^3 + \|x_i - x_{N_i}^*\|^3].$$

Hence, assuming, without loss of generality, that $K' \geq 1$ and that N_0 is sufficiently large to ensure that for all $N_i \geq N_0$, $\|x_{N_i}^* - x^*\| \leq 1$, we get

$$(4.12b) \quad \begin{aligned} \|x_{i+1} - x^*\|^2 &\leq 2[\|x_{i+1} - x_{N_i}^*\|^2 + \|x_{N_i}^* - x^*\|^2] \\ &\leq 2K'[\|x_{i+1} - x_i\|^3 + \|x_i - x_{N_i}^*\|^3 + \|x_{N_i}^* - x^*\|^2] \\ &\leq 8K'[\|x_{i+1} - x_i\|^3 + \|x_i - x^*\|^3 + \|x_{N_i}^* - x^*\|^3 + \|x_{N_i}^* - x^*\|^2] \\ &\leq 16K'[\|x_{i+1} - x_i\|^3 + \|x_i - x^*\|^3 + \|x_{N_i}^* - x^*\|^2]. \end{aligned}$$

Now, it follows from (4.4b) and (4.11d) that

$$(4.12c) \quad \|x_{N_i}^* - x^*\|^2 \leq \frac{2L}{mN_i} \leq \frac{2L}{m} \|x_{i+1} - x_i\|^3.$$

Substituting into (4.12b), we obtain that there exists a $K'' < \infty$, independent of N_i , such that

$$(4.12d) \quad \|x_{i+1} - x^*\|^2 \leq K''[\|x_{i+1} - x_i\|^3 + \|x_i - x^*\|^3].$$

To continue, let $B \triangleq \{x \in \mathbb{R}^n \mid \|x - x^*\| \leq 1\}$. It follows from Proposition 4.1 that there exists a $L' < \infty$ such that (4.4c), (4.4d) hold. As in the proof of Theorem 4.1, we will show that if N_0 is sufficiently large, then given any $\delta > 0$ there exists a $\rho > 0$ (where $\rho \leq 1$ without loss of generality) such that for all $N_i \geq N_0$ and $\|x_i - x^*\| \leq \rho$, $\|x_{i+1} - x_i\| \leq \delta$. Using the triangle inequality, (4.7d) and (4.4d), we obtain that

$$(4.13a) \quad \begin{aligned} \|x_{i+1} - x_i\| &= \|h_{N_i, e_i}(x_i)\| \\ &\leq \|h_{N_i, e_i}(x_i) - h_{N_i}(x_i)\| + \|h_{N_i}(x_i) - h(x_i)\| + \|h(x_i)\| \\ &\leq \sqrt{\frac{2\varepsilon_i}{m}} + \sqrt{\frac{4L'}{mN_i}} + \|h(x_i)\|. \end{aligned}$$

Furthermore, analogously to (4.8d), we have that

$$(4.13b) \quad \|x_{i+1} - x_i\| \leq \sqrt{\frac{|\theta_{N_i}(x_i)|}{m}} + \sqrt{\frac{4L'}{mN_i}} + \|h(x_i)\|.$$

Applying the triangle inequality once more and utilizing (4.4c), we obtain that

$$(4.13c) \quad \begin{aligned} \|x_{i+1} - x_i\| &\leq \sqrt{\frac{[|\theta(x_i)| + |\theta(x_i) - \theta_{N_i}(x_i)|]}{m}} + \sqrt{\frac{4L'}{mN_i}} + \|h(x_i)\| \\ &\leq \sqrt{\frac{[|\theta(x_i)| + L'/N_i]}{m}} + \sqrt{\frac{4L'}{mN_i}} + \|h(x_i)\|. \end{aligned}$$

It now follows from the continuity of $\theta(\cdot)$ and $h(\cdot)$ and the fact that $\theta(x^*) = 0$, $h(x^*) = 0$, that if N_0 is chosen sufficiently large and ρ sufficiently small, then the desired continuity result holds. We can now proceed as in the proof of Theorem 2.1, following (2.11b), to complete the proof. \square

5. Implementation of the global algorithm. To produce an implementation of Algorithm 3.1 (the global Newton method), we propose to use two mechanisms for controlling the precision of the approximations used. The first one will be taken from

Algorithm 4.2, and will ensure superlinear rate of convergence, while the second one, which we will allow to dominate the first one, will be an extension of the mechanism described in Appendix A of [23]. For our case, this extension can be described abstractly as follows. Suppose that for every integer $N \geq N_0 > 0$, $A_N : \mathbb{R}^n \rightarrow 2^{\mathbb{R}^n \times \mathbb{N}}$ (the set of all subsets of \mathbb{R}^n) is a (possibly) set-valued iteration map. The reason for introducing a second integer N' is that given an integer N , the algorithm may be required to increase it to a new value $N' \geq N$ before it can satisfy all the internal tests. Now consider the following algorithm model form solving the problem P in (2.1a).

Algorithm Model 5.1.

Data. $x_0 \in \mathbb{R}^n$, $N_0 \in \mathbb{N}$.

Step 0. Set $i = 0$.

Step 1. Set $N = N_i$.

Step 2. Compute a pair

$$(5.1a) \quad (y, N') \in A_N(x_i).$$

Step 3. If

$$(5.1b) \quad \psi_{N'}(y) - \psi_{N'}(x_i) \leq -\frac{1}{N'},$$

go to Step 4; else, replace N by $2N'$ and go to Step 2.

Step 4. Set $x_{i+1} = y$, $N_{i+1} = N$.

Step 5. Replace i by $i + 1$ and go to Step 1.

Our proof of convergence requires the following technical result.

LEMMA 5.1. *Suppose that the sequences of real numbers $\{\beta_i\}_{i=0}^\infty$ and $\{\eta_i\}_{i=0}^\infty$ satisfy the following conditions: (i) $\eta_i \geq 0$ for all $i \in \mathbb{N}$, (ii) $\sum_{i=0}^\infty \eta_i < \infty$, and (iii) $\beta_{i+1} \leq \beta_i + \eta_i$ for all $i \in \mathbb{N}$. Then either the sequence $\{\beta_i\}_{i=0}^\infty$ converges or $\beta_i \rightarrow -\infty$ as $i \rightarrow \infty$.*

Proof. It is clear from the assumptions that the following holds:

$$(5.2a) \quad \beta_n - \beta_0 = \sum_{i=0}^{n-1} (\beta_{i+1} - \beta_i) \leq \sum_{i=0}^\infty \eta_i.$$

Hence, β_i is bounded from above, and therefore $\hat{\beta} \triangleq \overline{\lim}_{i \rightarrow \infty} \beta_i < \infty$. Obviously, if $\hat{\beta} = -\infty$, then $\beta_i \rightarrow -\infty$ as $i \rightarrow \infty$.

Now suppose that $\hat{\beta} > -\infty$. To prove convergence of the sequence $\{\beta_i\}_{i=0}^\infty$, we will show by contradiction that $\lim_{i \rightarrow \infty} \beta_i = \hat{\beta}$. Thus let $\varepsilon > 0$ be arbitrary and suppose that there is no i_0 such that $\beta_i > \hat{\beta} - \varepsilon$ for all $i > i_0$. Clearly, there exists an i_1 such that $\sum_{k=i}^\infty \eta_k < \varepsilon/2$ for all $i \geq i_1$. It follows from our hypothesis that there exists an $i_2 \geq i_1$ such that $\beta_{i_2} \leq \hat{\beta} - \varepsilon$. It follows from (5.2a) that, for $i > i_2$,

$$(5.2b) \quad \beta_i - \beta_{i_2} = \sum_{k=i_2}^{i-1} (\beta_{k+1} - \beta_k) \leq \sum_{k=i_2}^\infty \eta_k \leq \frac{\varepsilon}{2}.$$

Hence $\beta_i \leq \hat{\beta} - \varepsilon/2$ for all i sufficiently large, which contradicts the definition of $\hat{\beta}$. It follows that $\lim_{i \rightarrow \infty} \beta_i = \hat{\beta}$. \square

THEOREM 5.1. *Suppose that Assumptions 2.1 and 4.1 hold, so that (4.4a) is valid, and that for every $x \in \mathbb{R}^n$ such that $0 \notin \partial\psi(x)$ there exist a $\rho_x > 0$, a $\delta_x > 0$, and an integer $N_x > 0$ such that*

$$(5.3) \quad \psi_{N'}(y') - \psi_{N'}(x') \leq -\delta_x$$

for all $N \geq N_x$, and all $x', y' \in \mathbb{R}^n$ such that $\|x' - x\| \leq \rho_x$, $(y', N') \in A_N(x')$.

Under these assumptions, if $\{x_i\}_{i=0}^\infty, \{N_i\}_{i=0}^\infty$ are a pair of sequences constructed by Algorithm Model 5.1, then $x_i \rightarrow x^*$ and $N_i \rightarrow \infty$ as $i \rightarrow \infty$, where x^* is the solution of (2.1a).

Proof. First, we use Lemma 5.1 to show that the sequence $\{\psi_{N_i}(x_i)\}_{i=0}^\infty$ converges. Let $I \triangleq \{i \in \mathbb{N} \mid N_{i+1} \neq N_i\}$, and let the sequence $\{\eta_i\}_{i=0}^\infty$ be defined by

$$(5.4a) \quad \eta_i \triangleq \begin{cases} L/N_i, & i \in I, \\ 0, & \text{otherwise,} \end{cases}$$

where L is the Lipschitz constant in Assumption 2.1(i). Now suppose that $i \in I$, and let $i_+ \triangleq \min \{j \in I \mid j > i\}$. Since by construction we have that $N_{i_+} \geq 2N_i$, it follows that

$$(5.4b) \quad \sum_{i=0}^\infty \eta_i = \sum_{i \in I} \eta_i \leq \sum_{k=0}^\infty \frac{L}{2^k N_0} < \infty.$$

Hence the sequence $\{\eta_i\}_{i=0}^\infty$ is summable. From (4.5c), we have that for all $i \in I$

$$(5.4c) \quad |\psi_{N_{i+1}}(x_i) - \psi_{N_i}(x_i)| \leq |\psi_{N_{i+1}}(x_i) - \psi(x_i)| + |\psi(x_i) - \psi_{N_i}(x_i)| \leq \frac{L}{2N_{i+1}} + \frac{L}{2N_i} \leq \frac{L}{N_i}.$$

Clearly, for all $i \in \mathbb{N} \setminus I$, i.e., such that $N_{i+1} = N_i$, $|\psi_{N_{i+1}}(x_i) - \psi_{N_i}(x_i)| = 0$. Hence for all $i \in \mathbb{N}$,

$$|\psi_{N_{i+1}}(x_i) - \psi_{N_i}(x_i)| \leq |\psi_{N_{i+1}}(x_i) - \psi(x_i)| + |\psi(x_i) - \psi_{N_i}(x_i)| \leq \frac{L}{2N_{i+1}} + \frac{L}{2N_i} \leq \frac{L}{N_i}.$$

Consequently, using (5.1b), we obtain that

$$(5.4d) \quad \begin{aligned} \psi_{N_{i+1}}(x_{i+1}) - \psi_{N_i}(x_i) &\leq \psi_{N_{i+1}}(x_{i+1}) - \psi_{N_{i+1}}(x_i) + \psi_{N_{i+1}}(x_i) - \psi_{N_i}(x_i) \\ &\leq -\frac{1}{N_{i+1}} + \eta_i \leq \eta_i. \end{aligned}$$

Furthermore, it follows from Proposition 2.1, (4.5c), and the fact that $N_{i+1} \geq N_i$ that

$$(5.4e) \quad \psi_{N_i}(x) \geq \psi(x^*) + \frac{m}{2} \|x - x^*\|^2 - \frac{L}{2N_0}.$$

Hence the sequence $\{\psi_{N_i}(x_i)\}_{i=0}^\infty$ satisfies the hypotheses of Lemma 5.1 and, in addition, it is bounded below. We therefore conclude that it converges.

Next, we show that $N_i \rightarrow \infty$. If this is not true, then since $N_{i+1} \geq N_i$ we must have $N_i = N^*$ for i sufficiently large. For such i , (5.1b) implies that

$$(5.4f) \quad \psi_{N^*}(x_{i+1}) - \psi_{N^*}(x_i) \leq -\frac{1}{N^*},$$

which contradicts the fact that the sequence $\{\psi_{N_i}(x_i)\}_{i=0}^\infty$ converges.

As a consequence of (5.4e), the sequence $\{x_i\}_{i=0}^\infty$ is bounded, and hence it must have accumulation points. For the sake of contradiction, suppose that the sequence $\{x_i\}_{i=0}^\infty$ does not converge to x^* . Then it must have an accumulation point $x^{**} \neq x^*$. Let $K \subset \mathbb{N}$ be the set of indices of the subsequence converging to x^{**} .

Since $x^{**} \neq x^*$, we have $0 \notin \partial\psi(x^{**})$, from which it follows by assumption that there exists a $\delta > 0$ such that for $i \in K$ sufficiently large,

$$(5.4g) \quad \psi_{N_{i+1}}(x_{i+1}) - \psi_{N_{i+1}}(x_i) \leq -\delta.$$

Referring to (5.4d), we see that for $i \in K$ sufficiently large,

$$(5.4h) \quad \psi_{N_{i+1}}(x_{i+1}) - \psi_{N_i}(x_i) \leq \psi_{N_{i+1}}(x_{i+1}) - \psi_{N_{i+1}}(x_i) + \psi_{N_{i+1}}(x_i) - \psi_{N_i}(x_i) \leq -\delta + \eta_i.$$

However, since $\eta_i \rightarrow 0$, (5.4h) contradicts the fact that the sequence $\{\psi_{N_i}(x_i)\}_{i=0}^\infty$ converges. Hence $0 \in \partial\psi(x^*)$. \square

The above algorithm model and our desire to retain the superlinear rate of convergence of the implementation of the local Newton method, Algorithm 4.2, leads us to the following algorithm.

Algorithm 5.2 (Implementable Global Newton Method for P).

Data. $x_0 \in \mathbb{R}^n$, $N_0 \in \mathbb{N}$, $\alpha, \beta \in (0, 1)$, $\varepsilon_0 > 0$, $K, \hat{\tau} \ll 1$, $S \triangleq \{1, \beta, \beta^2, \dots\}$.

Step 0. Set $i = 0$.

Step 1. Set $N = N_i$.

Step 2. Set $\varepsilon = \varepsilon_0$, $\tau = \hat{\tau}$.

Step 3. Compute a vector $h_{N,\varepsilon}(x_i) \in \mathbb{R}^n$ such that (see Comment 4.1)

$$(5.3a) \quad 0 \leq \hat{\psi}_N(x_i + h_{N,\varepsilon}(x_i) | x_i) - \psi_N(x_i) - \theta_N(x_i) \leq \varepsilon.$$

Step 4. If

$$(5.3b) \quad \hat{\psi}_N(x_i + h_{N,\varepsilon}(x_i)) - \psi_N(x_i) \leq -2\varepsilon$$

and

$$(5.3c) \quad \varepsilon \leq \|h_{N,\varepsilon}(x_i)\|^3,$$

go to Step 5.

Else,

if $\varepsilon \geq \tau$, replace ε by $\varepsilon/2$ and go to Step 3,

else, replace τ by $\tau/2$, N by $2N$, and go to Step 3.

Step 5. If

$$(5.3d) \quad \frac{K}{N} \leq \|h_{N,\varepsilon}(x_i)\|^3,$$

set $h_i = h_{N,\varepsilon}(x_i)$, and go to Step 6. Else, replace N by $2N$ and go to Step 3.

Step 6. Compute the stepsize

$$(5.3e) \quad \lambda_i \triangleq \max \{ \lambda \in S / \psi_N(x_i + \lambda h_i) - \psi_N(x_i) \leq \lambda \alpha [\hat{\psi}_N(x_i + h_i | x_i) - \psi_N(x_i)] \}.$$

Step 7. If

$$(5.3f) \quad \psi_N(x_i + \lambda h_i) - \psi_N(x_i) \leq -\frac{1}{N},$$

set $x_{i+1} = x_i + \lambda_i h_i$, $N_{i+1} = N$, and go to Step 8; else, replace N by $2N$ and go to Step 3.

Step 8. Replace i by $i+1$ and go to Step 1.

Theorem 5.1 can now be used to show that sequences constructed by Algorithm 5.2 converge to the solution x^* , while Theorem 4.2 leads to the conclusion that these sequences converge superlinearly.

THEOREM 5.2. *Suppose that Assumptions 2.1 and 4.1 hold and that x^* is the solution of problem (2.1a). Then any sequence $\{x_i\}_{i=0}^\infty$, constructed by Algorithm 5.2, converges to x^* superlinearly, with Q -rate at least $3/2$.*

Proof. The proof consists of two parts. The first part shows that the algorithm map $A_N(\cdot)$ satisfies the hypotheses of Theorem 5.1, from which we conclude that $x_i \rightarrow x^*$ as $i \rightarrow \infty$. The second part shows that for i sufficiently large, the stepsize λ_i is 1. In this case, Algorithm 5.2 reduces to the local Algorithm 4.2, and we may apply Theorem 4.2 (along with the fact that $N_i \rightarrow \infty$) to conclude that the iterates converge superlinearly.

(a) To show that $A_N(\cdot)$ satisfies the hypotheses of Theorem 5.1, suppose that $x \in \mathbb{R}^n$ is such that $0 \notin \psi(x)$. Then Lemma 2.2(d) implies that $\theta(x) < 0$. By continuity, there exist $\rho > 0$, $\delta > 0$ such that for all $x' \in B(x, \rho) \triangleq \{x' \in \mathbb{R}^n \mid \|x' - x\| \leq \rho\}$, $\theta(x') \leq -\delta < 0$. Furthermore, since $B(x, \rho)$ is bounded, (4.4c) implies that there exists an $\bar{N} \in \mathbb{N}$ such that for all $N \geq \bar{N}$ and $x' \in B(x, \rho)$ $\theta_N(x') \leq -\delta/2$.

Suppose that $x' \in B(x, \rho)$ and that the algorithm map $A_N(\cdot)$ produces a pair $(y', N') \in A_N(x')$ (with $N' \geq N$), and an ε satisfying the tests in Steps 3 and 4. Then from (5.3b) we have

$$(5.5) \quad \theta_{N'}(x') \leq \hat{\psi}_{N'}(x' + h_{N',\varepsilon}(x') \mid x') - \psi_{N'}(x') \leq -2\varepsilon,$$

which yields $-\theta_{N'}(x')/2 \geq \varepsilon$. Furthermore, using (5.3a), we obtain that

$$(5.6) \quad \hat{\psi}_{N'}(x' + h_{N',\varepsilon}(x') \mid x') - \psi_{N'}(x') \leq \varepsilon + \theta_{N'}(x') \leq \frac{\theta_{N'}(x')}{2}.$$

It follows from the convexity of the function $\lambda \mapsto \hat{\psi}_{N'}(x' + \lambda h_{N',\varepsilon}(x') \mid x') - \psi_{N'}(x')$ that for all $\lambda \in [0, 1]$,

$$(5.7) \quad \hat{\psi}_{N'}(x' + \lambda h_{N',\varepsilon}(x') \mid x') - \psi_{N'}(x') \leq \lambda [\hat{\psi}_{N'}(x' + h_{N',\varepsilon}(x') \mid x') - \psi_{N'}(x')].$$

By Lemma 2.1, applied to the functions $\psi_{N'}$, $\hat{\psi}_{N'}$, we have the estimate

$$(5.8) \quad |\hat{\psi}_{N'}(x' + \lambda h_{N',\varepsilon}(x') \mid x') - \psi_{N'}(x' + \lambda h_{N',\varepsilon}(x'))| \leq \hat{K} \lambda^3 \|h_{N',\varepsilon}(x')\|^3.$$

Combining (5.7) and (5.8) yields

$$(5.9) \quad \begin{aligned} & \psi_{N'}(x' + \lambda h_{N',\varepsilon}(x')) - \psi_{N'}(x') \\ & \leq \hat{K} \lambda^3 \|h_{N',\varepsilon}(x')\|^3 + \lambda [\hat{\psi}_{N'}(x' + h_{N',\varepsilon}(x') \mid x') - \psi_{N'}(x')]. \end{aligned}$$

Using (5.6) and (5.9), we obtain that for all $\lambda \in [0, 1]$,

$$(5.10) \quad \begin{aligned} & \psi_{N'}(x' + \lambda h_{N',\varepsilon}(x')) - \psi_{N'}(x') - \alpha \lambda [\hat{\psi}_{N'}(x' + h_{N',\varepsilon}(x') \mid x') - \psi_{N'}(x')] \\ & \leq \hat{K} \lambda^3 \|h_{N',\varepsilon}(x')\|^3 + (1 - \alpha) \lambda [\hat{\psi}_{N'}(x' + h_{N',\varepsilon}(x') \mid x') - \psi_{N'}(x')] \\ & = \lambda ((1 - \alpha) [\hat{\psi}_{N'}(x' + h_{N',\varepsilon}(x') \mid x') - \psi_{N'}(x')] + \hat{K} \lambda^2 \|h_{N',\varepsilon}(x')\|^3) \\ & \leq \lambda \left((1 - \alpha) \frac{\theta_{N'}(x')}{2} + \hat{K} \lambda^2 \|h_{N',\varepsilon}(x')\|^3 \right) \\ & \leq \lambda \left(-(1 - \alpha) \frac{\delta}{4} + \hat{K} \lambda^2 \|h_{N',\varepsilon}(x')\|^3 \right). \end{aligned}$$

Combining the facts that $h(\cdot)$ is continuous and $B(x, \rho)$ is bounded with (4.4d) and (4.7d), we conclude that there exists a constant $\Delta < \infty$ such that for all $x' \in B(x, \rho)$ and all $N' \geq N$, $\|h_{N',\varepsilon}(x')\| \leq \Delta$. Thus (5.10) yields

$$(5.11) \quad \begin{aligned} & \psi_{N'}(x' + \lambda h_{N',\varepsilon}(x')) - \psi_{N'}(x') - \alpha \lambda [\hat{\psi}_{N'}(x' + h_{N',\varepsilon}(x') \mid x') - \psi_{N'}(x')] \\ & \leq \lambda \left(-(1 - \alpha) \frac{\delta}{4} + \hat{K} \lambda^2 \Delta^3 \right), \end{aligned}$$

from which we conclude that there exists a $0 < \lambda_0 < 1$ such that for all $x' \in B(x, \rho)$, the stepsize λ' produced by Step 6 satisfies $\lambda' \geq \lambda_0$. Consequently, using (5.3e), (5.6), and the fact that $\theta_N(x') \leq -\delta/2$, we obtain that for all $x' \in B(x, \rho)$ and all $N' \geq N \geq \bar{N}$,

$$(5.12) \quad \begin{aligned} & \psi_{N'}(x' + \lambda' h_{N',\varepsilon}(x')) - \psi_{N'}(x') \leq \alpha \lambda_0 [\hat{\psi}_{N'}(x' + h_{N',\varepsilon}(x') \mid x') - \psi_{N'}(x')] \\ & \leq -\frac{\alpha \lambda_0 \delta}{4}. \end{aligned}$$

If we let $N_x = \bar{N}$, $\rho_x = \rho$, and $\delta_x = \alpha\lambda_0\delta/4$, we see that the map $A_N(\cdot)$ satisfies the conditions of Theorem 5.1. Hence $x_i \rightarrow x^*$, and as a consequence of Step 7, we see that $N_i \rightarrow \infty$.

(b) To complete the proof, we must show that for i sufficiently large, the stepsize λ_i is 1. If we set $\lambda = 1$ in the second to last line of (5.10) we obtain that

$$(5.13) \quad \begin{aligned} & \psi_{N_{i+1}}(x_i + h_i) - \psi_{N_{i+1}}(x_i) - \alpha[\hat{\psi}_{N_{i+1}}(x_i + h_i | x_i) - \psi_{N_{i+1}}(x_i)] \\ & \leq \left((1 - \alpha) \frac{\theta_{N_{i+1}}(x_i)}{2} + \hat{K} \|h_i\|^3 \right), \end{aligned}$$

where h_i is as defined in Step 5 of Algorithm 5.2. Using a result similar to (3.4) we obtain that $\theta_{N_{i+1}}(x_i) \leq -(m/2)\|h_{N_{i+1}}(x_i)\|^2$. Using (4.7d), (5.3c), and the fact that $\|x\|^2 \geq \frac{1}{2}\|y\|^2 - \|x - y\|^2$, we obtain that

$$(5.14) \quad \|h_{N_{i+1}}(x_i)\|^2 \geq \frac{1}{2}\|h_i\|^2 - \|h_i - h_{N_{i+1}}(x_i)\|^2 \geq \frac{1}{2}\|h_i\|^2 - \frac{2\varepsilon_i}{m} \geq \frac{1}{2}\|h_i\|^2 - \frac{2}{m}\|h_i\|^3.$$

Hence we obtain the bound

$$(5.15) \quad \theta_{N_{i+1}}(x_i) \leq \frac{-m}{4}\|h_i\|^2 + \|h_i\|^3.$$

Substituting this bound into (5.13), we obtain that

$$(5.16) \quad \begin{aligned} & \psi_{N_{i+1}}(x_i + h_i) - \psi_{N_{i+1}}(x_i) - \alpha[\hat{\psi}_{N_{i+1}}(x_i + h_i | x_i) - \psi_{N_{i+1}}(x_i)] \\ & \leq \frac{(1 - \alpha)}{2} \left(\frac{-m}{4}\|h_i\|^2 + \|h_i\|^3 \right) + \hat{K}\|h_i\|^3. \end{aligned}$$

From (4.13b) we note that $h_i \rightarrow 0$ as $i \rightarrow \infty$, and hence the right-hand side of (5.16) is negative for i sufficiently large. Hence $\lambda_i = 1$ for i sufficiently large. This completes the proof. \square

6. A numerical example. We will present the solution of a semi-infinite minimax problem that was constructed by converting an optimal control problem with control and state-space constraints, by means of an exact penalty function, into an unconstrained minimax problem.

The original optimal control problem is as follows:

$$(6.1) \quad \min_{x \in \mathbb{R}^{21}} \left\{ \frac{1}{2}(\|z(x, 1)\|^2 + 10^{-6}\|x\|^2) \mid z^2(x, t) - 0.15 \leq 0, \forall t \in [0, 20], x_j^2 - 1 \leq 0, \forall j \in \mathbf{p} \right\},$$

where $\mathbf{p} \triangleq \{0, \dots, 20\}$, and $z: \mathbb{R}^{21} \times [0, 20] \rightarrow \mathbb{R}^2$ solves the differential equation

$$(6.2) \quad \frac{d}{dt} z(x, t) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} z(x, t) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(x, t),$$

where $x(z, 0) = (-2.5, 0)^T$, and the control $u: \mathbb{R}^{21} \times [0, 20] \rightarrow \mathbb{R}$ is defined by the 21-dimensional parameter vector x , through linear interpolation, as follows: for any $j = 0, 1, \dots, 19$, and $t = \lambda j + (1 - \lambda)(j + 1)$, $u(x, j) = \lambda x_j + (1 - \lambda)x_{j+1}$. Note that the dynamics in (6.1) are so simple that we can integrate them exactly for the resulting piecewise linear control.

To convert the above optimal control problem into unconstrained form, we use a parameter of 100 in the exact penalty function, and obtain a problem of the form (2.1a), (2.1b), with $q = 23$, and the functions $\phi^j(\cdot, \cdot)$, $j \in \mathbf{q}$ defined by

$$(6.3a) \quad \phi^1(x, t) \triangleq \frac{1}{2}(\|z(x, 1)\|^2 + 10^{-6}\|x\|^2),$$

$$(6.3b) \quad \phi^2(x, t) \triangleq \phi^1(x, t) + 100(x_2(x, t/20) - 0.15),$$

$$(6.3c) \quad \phi^j(x, t) \triangleq \phi^1(x, t) + 100(x_{j-2}^2 - 1), \quad \forall j \in \{3, \dots, 23\},$$

with $t \in [0, 1]$. The algorithm was started from the initial point $x_0 = (1, -1, 1, -1, 1, -1, 1, -1, 1, -1, 1, -1, 1, -1, 1, -1, 1)$.

Figures 1 and 2 present our computational results. Figure 1 presents plots of the control sequence x^i versus i , while Fig. 2 presents phase-plane plots, i.e., plots of $x^1(t)$ versus $x^2(t)$, traditional in the control literature. We see from these plots that problem (6.1) is solved in two iterations, at the end of which the original semi-infinite constraint, $z^2(u, t) - 0.15 \leq 0$ for all $t \in [0, 20]$ has been satisfied. The initial value of the cost was $\psi_5(x_0) = 3.12501$, the final value of the cost was $\psi_{28}(2) = 2.09003 \times 10^{-7}$; the final value of the optimality function was $\theta_{28}(x_2) = -2.01827 \times 10^{-7}$. To limit the growth of the discretization index, we set $K = 10^{-20}$ in (5.3d) so as to be able to satisfy (5.3d) with $N = 100$ when $\|h_{N,\varepsilon}\| 10^{-10}$. We set $N_0 = 5$, and the algorithm set $N_1, N_2 = 28$.

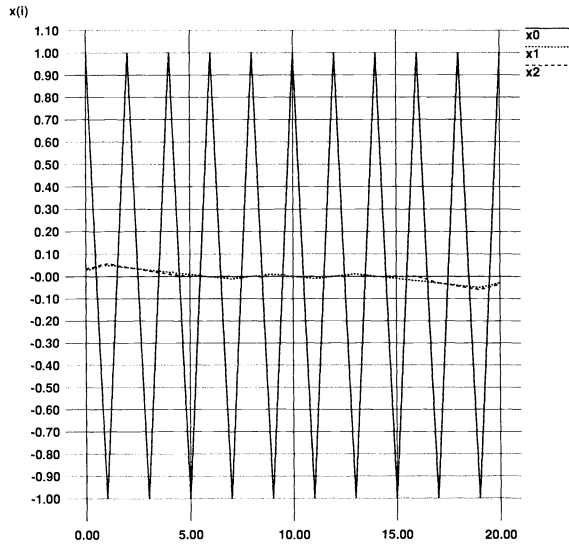


FIG. 1. Plots of control sequence at iterations 0, 1, 2.

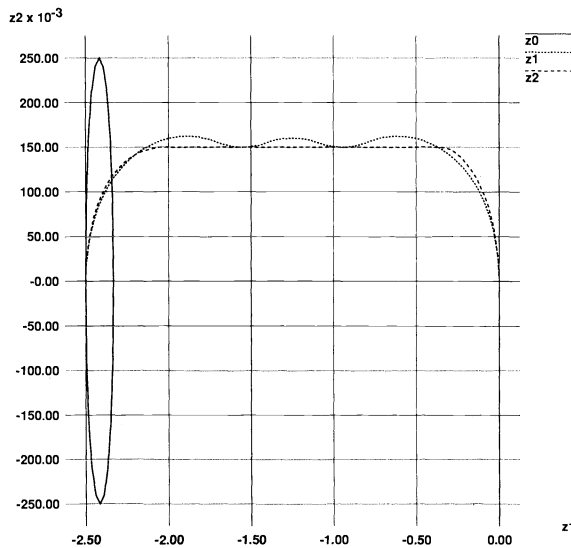


FIG. 2. Phase-plane plots at iterations 0, 1, 2.

To reduce the looping in the computation of the discretization parameter N , instead of simply doubling it whenever the test (5.3d) failed to be satisfied, we set the next value to the integer part of $\max \{6.4 \times 10^{-19} / \|h_{N,\epsilon}\|^3, 1.1N\}$.

We compute the search direction $h_{N,\epsilon}(x)$, satisfying (5.3a), by applying to the dual of problem (4.2e) the rapidly converging Levitin–Polyak constrained Newton method [13].

To illustrate how inequality (5.3a) is solved, consider the primal problem (4.2e). When expanded, it assumes the form

$$(6.4a) \quad \theta \triangleq \min_{h \in \mathbb{R}^n} \max_{j \in \mathbf{q}} f^j + \langle g_j, h \rangle + \frac{1}{2} \langle h, H^j h \rangle.$$

The dual of this problem is

$$(6.4b) \quad \theta = -\min_{\mu \in \Sigma} J_d(\mu),$$

where $J_d(\mu) \triangleq -\sum_{j=1}^q \mu^j f^j + \frac{1}{2} \langle \sum_{j=1}^q \mu^j g_j, (\sum_{j=1}^q \mu^j H^j)^{-1} \sum_{j=1}^q \mu^j g_j \rangle$, and Σ is the unit simplex in \mathbb{R}^n . The formula for the second derivative matrix of the dual cost function, required by the Levitin–Polyak method, is given in an appendix in [26].

When applied to the dual problem, the Levitin–Polyak method computes a sequence $\{\mu_i\}_{i=0}^\infty$, which converges quadratically to \hat{h} , a solution of the dual problem (6.4b). Furthermore, if we define $h: \Sigma \rightarrow \mathbb{R}^n$ by

$$(6.5) \quad h(\mu) \triangleq \left(\sum_{j=1}^q \mu^j H^j \right)^{-1} \sum_{j=1}^q \mu^j g_j,$$

it is easy to show that the corresponding sequence $\{h(\mu_i)\}_{i=0}^\infty$ converges to \hat{h} , the unique solution of the primal problem (6.4a). Noting that the iterations of the Levitin–Polyak algorithm generate both upper and lower bounds on θ , as given by

$$(6.6a) \quad J_p(\mu_i) \geq \theta \geq -J_d(\mu_i),$$

where

$$(6.6b) \quad J_p(\mu_i) \triangleq \max_{j \in \mathbf{q}} f^j + \langle g_j, h(\mu_i) \rangle + \frac{1}{2} \langle h(\mu_i), H^j h(\mu_i) \rangle,$$

and making use of the fact that $J_p(\mu_i) - J_d(\mu_i) \rightarrow 0$, we see that a point $h(\mu_i)$ satisfying (5.3a) can be computed in a finite number of iterations of the Levitin–Polyak method.

7. Conclusion. We have used a new and very simple proof technique to show that natural *conceptual* extensions of Newton's method converge superlinearly on a class of semi-infinite minimax problems. This technique has also enabled us to construct rate-preserving *implementations* of these extensions. Our implementations are interesting for two reasons: first, they account for all the significant approximations involved, and second, they do not require the knowledge of the Lipschitz constants or eigenvalue bounds associated with the problem functions and their first- and second-order derivatives.

Apart from the intrinsic interest that a theoretical extension of Newton's method to semi-infinite optimization possesses, our numerical results show that it is a viable procedure for the solution of such classical problems as state- and control-constrained optimal control problems with linear dynamics.

Acknowledgments. We thank the referees for their most helpful comments.

REFERENCES

- [1] L. ARMIJO, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific J. Math., 16 (1966), pp. 1-3.
- [2] C. BERGE, *Topological Spaces*, Macmillan, New York, 1963.
- [3] S. P. BOYD AND H. B. CRAIG, *Linear Control Design, Limits of Performance*, Prentice-Hall, Inform. System Sci. Ser., Englewood Cliffs, NJ, 1991.
- [4] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.
- [5] I. D. COOPE AND G. A. WATSON, *A projected Lagrangian algorithm for semi-infinite programming*, Math. Programming, 32 (1985), pp. 337-356.
- [6] J. E. DENNIS, JR. AND R. B. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.
- [7] J. DIEUDONNE, *Foundations of Modern Analysis*, Academic Press, New York, 1960.
- [8] A. A. GOLDSTEIN, *Constructive Real Analysis*, Harper and Row, New York, 1967.
- [9] S. P. HAN, *Variable metric methods for minimizing a class of nondifferentiable functions*, Math. Programming, 20 (1981), pp. 1-13.
- [10] L. HE AND E. POLAK, *Effective diagonalization strategy for the solution of a class of optimal design problems*, IEEE Trans. Automat. Control, 35 (1990), pp. 258-267.
- [11] R. HETTICH AND W. VAN HOSSEDE, *On quadratically convergent methods for semi-infinite programming*, in *Semi-infinite Programming*, Lecture Notes in Control and Inform. Sci. 15, R. Hettich, ed., Springer-Verlag, New York, 1979, pp. 97-111.
- [12] R. KLESSIG AND E. POLAK, *An adaptive algorithm for unconstrained optimization with applications to optimal control*, SIAM J. Control, 11 (1973), pp. 80-94.
- [13] E. S. LEVITIN AND B. T. POLYAK, *Constrained minimization methods*, U.S.S.R. Comput. Math. and Math. Phys., 6 (1966), pp. 1-50.
- [14] O. L. MANGASARIAN, *Nonlinear Programming*, McGraw-Hill, New York, 1969.
- [15] D. Q. MAYNE, *On the use of exact penalty functions to determine the step length of optimization algorithms*, Lecture Notes in Math., 773, Numerical Analysis, Springer-Verlag, Berlin, New York, 1980, pp. 98-109.
- [16] D. Q. MAYNE AND E. POLAK, *A quadratically convergent algorithm for solving infinite dimensional inequalities*, Appl. Math. Optim., 9 (1982), pp. 25-40.
- [17] ———, *A superlinearly convergent algorithm for constrained optimization problems*, Math. Programming Study, 16 (1982), pp. 45-61.
- [18] H. MUKAI AND E. POLAK, *A second order algorithm for unconstrained optimization*, J. Optim. Theory Appl., 26 (1978), pp. 501-513.
- [19] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
- [20] M. L. OVERTON, *On minimizing the maximum eigenvalue of a symmetric matrix*, in *Linear Algebra in Signals, Systems and Control*, B. N. Datta, ed., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1988.
- [21] E. POLAK, R. W. H. SARGENT, AND D. J. SEBASTIAN, *On the convergence of sequential minimization algorithms*, J. Optim. Theory Appl., 14 (1974), pp. 439-442.
- [22] E. POLAK, *On the mathematical foundations of nondifferentiable optimization in engineering design*, SIAM Rev., 29 (1987), pp. 21-89.
- [23] ———, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, 1971.
- [24] E. POLAK AND A. L. TITS, *A recursive quadratic programming algorithm for semi-infinite optimization problems*, Appl. Math. Optim., 8 (1982), pp. 325-349.
- [25] E. POLAK, *A modified secant method for unconstrained optimization*, Math. Programming, 6 (1974), pp. 264-280.
- [26] E. POLAK, D. Q. MAYNE, AND J. HIGGINS, *A superlinearly convergent algorithm for minimax problems*, University of California, Berkeley, Electronics Research Laboratory Memo No. M86/103, November 1986. J. Optim. Theory Appl., to appear.
- [27] E. POLAK, J. HIGGINS, AND D. Q. MAYNE, *A barrier function method for minimax problems*, University of California, Electronics Research Laboratory, Memo No. UCB/ERL M88/64, October 1988. Math. Programming, to appear.

- [28] E. POLAK AND D. Q. MAYNE, *An exact penalty function algorithm for control problems with state and control constraints*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 380-387.
- [29] S. M. ROBINSON, *Perturbed Kuhn-Tucker points and rates of convergence for a class of nonlinear-programming algorithms*, Math. Programming, 7 (1974), pp. 1-16.
- [30] ———, *Extension of Newton's method to mixed systems of nonlinear equations and inequalities*, Numer. Math., 19 (1972), pp. 341-347.
- [31] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

A PROBLEM DECOMPOSITION TECHNIQUE WITH APPLICATION TO THE OPTIMAL DISTRIBUTION OF ENZYMES*

DAVID E. STEWART†

Abstract. A model for enzyme-driven reactions in steady state is given which involves the enzyme concentrations linearly and the reactant concentrations nonlinearly. From this model an optimization problem is set up for the enzyme distributions, which involves only nonnegativity and integral constraints. Using some earlier theoretical work of Holm aker and Stewart [*SIAM J. Control Optim.* 25 (1987), pp. 1032–1052], a problem decomposition principle is proven. Results for a single enzyme-driven reaction is then used to produce a dynamic programming technique for trees and chains of enzyme driven reactions. This is extended to give a technique for determining the optimal amount of each enzyme subject to an overall cost limit. Numerical results are given for a test problem of the author's devising.

Key words. enzyme kinetics, optimal control, decomposition techniques

AMS(MOS) subject classifications. 80A30, 49B10, 49C20

1. Introduction. In this paper we consider a class of optimisation problems related to enzyme-driven reactions in one-dimensional continuous flow reactors. This model seems to be appropriate for modelling reactions in liver tubules [2] and for “packed bed” reactors, where a solution containing reactants flows past enzymes (which are held fixed) that act on the reactants. Assuming that the system has negligible diffusion *along* the liver tubule or packed bed compared to convection, and diffusion dominates *across* the tubule or packed bed, which is in a steady state, then each reactant concentration satisfies a first-order *ODE*. It is assumed that the enzyme concentrations appear linearly in the ODEs, but that each reactant concentration may appear nonlinearly. The enzyme concentrations are considered to be the control functions and are subject only to nonnegativity and integral constraints. That is, negative concentrations are not allowed, and a fixed amount of each enzyme is available. If a variable amount is available, then there is usually a cost for the enzyme, and by analysing first the case where the amount of enzyme is fixed, the case where the amount of enzyme available is variable can be analysed as is done in §7. The crucial point is that there is no *a priori* bound on the concentrations of enzymes or on the mixtures of enzymes allowed.

The first such problem to appear in the literature to the author's knowledge is due to Bass, Bracken, and Vyborny[2], which deals with a problem involving two enzymes and is of interest in relation to enzyme distribution in the liver. In [2], the following model for the liver's elimination of a drug Phenacitin was used. Justification for this model can be found in [1]. Diagrammatically, the elimination proceeded in two steps $P \rightarrow M \rightarrow \hat{M}$, where P represents Phenacitin, M a metabolite, and \hat{M} a “conjugated” metabolite which could be safely eliminated from the body. We also use $P(x)$ to denote the *concentration* of Phenacitin at a point at distance x along a liver tubule. Similarly for $M(x)$ and $\hat{M}(x)$. If we let $f(x)$, $g(x)$ denote the concentrations of the first and second enzymes and $\alpha(P(x))$ denote the reaction rate of the first reaction step per

*Received by the editors April 25, 1990; accepted for publication (in revised form) February 18, 1991.

† University of Queensland, Department of Mathematics, St. Lucia, Queensland 4072, Australia.

unit enzyme, and $\beta(M(x))$ denote the reaction rate of the second reaction step per unit enzyme, then the differential equations for this system are:

$$\begin{aligned} P' &= dP/dx = -f(x)\alpha(P(x)), & P(0) &= P_0 \\ M' &= dM/dx = +f(x)\alpha(P(x)) - g(x)\beta(M(x)), & M(0) &= 0 \\ \hat{M}' &= d\hat{M}/dx = & +g(x)\beta(M(x)), & \hat{M}(0) = 0. \end{aligned}$$

The issue in [2] was that the metabolite M is toxic to the body, and therefore the output from the liver should contain as little of it as possible, while still maintaining the overall ability of the body to eliminate Phenacitin. This essentially means that the amount of conjugated metabolite \hat{M} in the output should be maximized, given the amount of each enzyme to be used.

The question of whether the this optimization is actually performed in nature is a matter for experimental verification, although it is not uncommon in theoretical biology to assume that over time nature finds "solutions" that are nearly optimal for the problems that organisms face. Of course, if the system under study (e.g., a packed bed reactor) is designed by humans, then it is the duty of the relevant engineer(s) to ensure that it operates as close to optimal as is possible.

Bass, Bracken, and Vyborny [2] used an integral equation approach to obtain results for certain cases. Fink [5] used a similar approach to obtain more complete results, while Holm aker [6] also obtained further results, but used Pontryagin's maximum principle instead. Holm aker and Stewart [7] used an invariance result to reduce a much larger class of such problems to a problem with (pointwise) bounded controls; they then used this approach to solve a problem involving three enzymes. The class of optimal control problems considered in [7] is to realize

$$(1.1a) \quad \min_{y,u} C(y,u) \quad \text{where } C(y,u) = g(y(L)) + \int_0^L d(y(x))^T u(x) dx$$

subject to

$$(1.1b) \quad \begin{aligned} y' &= dy/dx = A(y)u, & y(0) &\in K; & K &\text{compact} \\ u(x) &\geq 0 \quad \text{for all } x \in [0, L], & \int_0^T u(x) dx &= a \geq 0 \end{aligned}$$

where $y \in C([0, L], \mathbf{R}^n)$ and $u \in L^1([0, L], \mathbf{R}^m)$. Here $u_j(x)$ is to be interpreted to be the concentration of enzyme j , and $y_i(x)$ is the concentration of reactant i , at distance x from the point of inflow of material into the reactor or liver.

In this paper we wish to extend the invariance result of Holm aker and Stewart [7], and to show how this extension can be used to decompose (and thus solve) a class of large-scale optimal enzyme distribution problems. The key to decomposing such problems is to split the system into two parts, the first of which is independent of the second, and the second depends on the first only through a single scalar function of a suitable form. If the optimization criterion only directly involves the second subsystem, then the connection between the two subsystems can be reduced to a single number, which is the integral of the input to the second subsystem. In relation to enzyme driven reactions, this means that the problem of optimizing (say) the chain of reactions $P \rightarrow M \rightarrow \hat{M}$ can be reduced to that of maximizing the output of $P \rightarrow M$, and then maximizing the output of $\theta \rightarrow M \rightarrow \hat{M}$ where the θ is the source of M ; it is the *total* input from θ that is important, not how it is done.

More precisely, we assume that the optimal control problem has the following structure. Realize

$$(1.2a) \quad \min_{y_1, y_2, u_1, u_2} C(y_2, u_2) \quad \text{where } C(y_2, u_2) = g(y_2(L)) + \int_0^L d(y_2(x))^T u_2(x) dx$$

subject to

$$(1.2b) \quad \begin{aligned} S_1: y_1' &= A_1(y_1) u_1, & y_1(0) &\in K_1 \subset \mathbf{R}^{n_1} \\ u_1(x) &\geq 0, & \int_0^L u_1(x) dx &= a_1 \geq 0 \in \mathbf{R}^{m_1} \\ S_2: y_2' &= A_2(y_2) u_2 + c(y_1)^T u_1 \cdot b(y_2), & y_2(0) &\in K_2 \subset \mathbf{R}^{n_2} \\ u_2(x) &\geq 0, & \int_0^L u_2(x) dx &= a_2 \geq 0 \in \mathbf{R}^{m_2} \end{aligned}$$

where K_1, K_2 are compact and $c(y_1) \geq 0$ for all $y_1 \in \mathbf{R}^{n_1}$. The single scalar function that is treated as input to the second subsystem is $\theta(x) = c(y_1(x))^T u_1(x)$ which we assume to be nonnegative. Even if only $\lambda = \int_0^L \theta(x) dx$ is known, the infimum of $C(y_2, u_2)$ can be computed. Thus the optimization problems for the two subsystems can then (almost) be treated in isolation, and the results can be recombined by means of an extension of the invariance result in [7].

If the system of enzyme-driven reactions forms a chain or a tree, then the above decomposition techniques can be applied repeatedly until we only consider one-enzyme systems, possibly with an external input. The biologically important case of an enzyme-driven *cycle* is not decomposable by this scheme. The analytical results for one-enzymes systems that are known (see [2], [5], [7]) are extended and used to develop numerical methods for solving the optimal distribution problem for trees of reactions by means of dynamic programming. These results and methods can be extended to deal with the problem of finding the optimal *amount* of each enzyme, as well as the optimal distribution, according to some total enzyme cost criterion. These numerical methods have been implemented, and the numerical results obtained for a test problem are presented.

2. Previous theoretical results. In this section we wish to review the theoretical results of Holm aker and Stewart [7] that are of relevance here. Holm aker and Stewart [7] consider optimal control problems of the form

$$(2.1a) \quad \min_{y, u} C(y, u) \quad \text{where } C(y, u) = g(y(L)) + \int_0^L d(y(x))^T u(x) dx$$

subject to

$$(2.1b) \quad \begin{aligned} y' &= dy/dx = A(y) u, & y(0) &\in K; & K &\text{compact} \\ u(x) &\geq 0 \quad \text{for all } x \in [0, L], & \int_0^T u(x) dx &= a \geq 0 \end{aligned}$$

where $y \in C([0, L], \mathbf{R}^n)$ and $u \in L^1([0, L], \mathbf{R}^m)$. We make the following assumptions throughout the remainder of the paper:

- (1) $A: \mathbf{R}^n \rightarrow \mathbf{R}^{n \times m}$ is Lipschitz continuous.
- (2) K is compact and path connected.
- (3) g and d are continuous functions.

In [7], assumption (1) was weakened to simply requiring that A be continuous and satisfy a boundedness condition of the form

$$y^T A(y) u \leq C_0(1 + \|y\|^2) \|u\| \quad \text{for all } y \in \mathbf{R}^n \text{ and } u \in (\mathbf{R}_+)^m.$$

Assumption (2) can also be weakened to only requiring that K is compact and has only finitely many path-connected components. Then each path-connected component can be handled separately.

A major difficulty in applying the standard qualitative optimal control theory to (2.1) is that for each x the set of admissible control values $u(x)$ is not compact. To get around this difficulty, Holm aker and Stewart [7] showed that (2.1) is essentially invariant under the following transformation, which is called a *time transform* in [7].

DEFINITION 2.1. For $u, v \geq 0$, $u, v \in L^1(0, L)^m$ and absolutely continuous functions y, z on $[0, L]$, (v, z) is said to be an x -transform of (u, y) if there is an absolutely continuous (AC) nondecreasing surjective function $\phi: [0, L] \rightarrow [0, L]$ such that

$$(2.3a) \quad v(x) = u(\phi(x)) \cdot \phi'(x) \quad \text{for a.a. } x \in [0, L]$$

$$(2.3b) \quad z(x) = y(\phi(x)) \quad \text{for all } x \in [0, L].$$

Furthermore, we say that v is an x -transform of u if (2.3a) holds.

The following results were then proven in [7]:

LEMMA 2.2. If (v, z) is an x -transform of (u, y) and (u, y) satisfies (2.1b) then so does (v, z) , and further, $C(u, y) = C(v, z)$.

Proof. See [7, Lem. 2.1]. \square

The next result shows that we do not need to look at controls that are outside a weakly compact set.

LEMMA 2.3. If (v, z) satisfies (2.1b) then (v, z) is a time transform of some (u, y) that satisfies (2.1b) and

$$(2.4) \quad \sum_{i=1}^m u_i(x) = L^{-1} \sum_{i=1}^m a_i \quad \text{for all } x.$$

Proof. See [7, Lem. 2.2]. \square

Combining these results gives Theorem 2.4.

THEOREM 2.4. Solutions exist for problem (2.1).

Proof. A detailed proof is given in [6, Thm. 2.1]. The basic idea is sketched here.

By Lemma 2.3 and Lemma 2.2 it is sufficient to restrict attention to the set defined by (2.4) and $u(x) \geq 0$, which is closed and bounded in $L^1[0, L]$. By Alaoglu's theorem, this set is weak* compact in $L^1[0, L]$. As the functional defined by (2.1) is weak* continuous, it follows that there must be a minimum. \square

3. Problems with input. Consider the problem of realizing

$$(1.2a) \quad \min_{y_1, y_2, u_1, u_2} C(y_2, u_2) \quad \text{where } C(y_2, u_2) = g(y_2(L)) + \int_0^L d(y_2(x))^T u_2(x) dx$$

subject to

$$(1.2b) \quad \begin{aligned} S_1: y_1' &= A_1(y_1) u_1, & y_1(0) &\in K_1 \subset \mathbf{R}^{n_1} \\ u_1(x) &\geq 0, & \int_0^L u_1(x) dx &= a_1 \geq 0 \in \mathbf{R}^{m_1} \\ S_2: y_2' &= A_2(y_2) u_2 + c(y_1)^T u_1 \cdot b(y_2), & y_2(0) &\in K_2 \subset \mathbf{R}^{n_2} \\ u_2(x) &\geq 0, & \int_0^L u_2(x) dx &= a_2 \geq 0 \in \mathbf{R}^{m_2} \end{aligned}$$

where K_1, K_2 are compact and $c(y_1) \geq 0$ for all $y_1 \in \mathbf{R}^{n_1}$. This system can be split into two parts, S_1 and S_2 , and here we concentrate on S_2 . Let $\theta(x) = c(y_1(x))^T u_1(x)$,

so that under the assumption that $c(y_1) \geq 0$ for all $y_1 \in \mathbf{R}^{n_1}$, $\theta(x) \geq 0$ for all $x \in [0, L]$ and $\theta \in L^1[0, L]$. Then given optimal y_1 and u_1 , (1.2) is equivalent to the problem of realizing

$$(3.1a) \quad \min_{y_2, u_2} C(u_2, y_2)$$

subject to

$$(3.1b) \quad \begin{aligned} y_2' &= A_2(y_2) u_2 + b(y_2) \theta(x), & y_2(0) &\in K_2 \\ u_2(x) &\geq 0, & \int_0^L u_2(x) dx &= a_2 \geq 0. \end{aligned}$$

The subscripts will be dropped for the remainder of this section as it will be understood that we are working with subsystem S_2 . As θ is fixed (by subsystem S_1), the results of the previous section are not applicable and the minimum may not be attainable. However, replacing the minimum by an infimum does give a well-posed problem. Let

$$C^*(\theta) = \inf \{ C(u, y) \mid (u, y) \text{ satisfy (3.1)} \}.$$

THEOREM 3.1. *$C^*(\theta)$ is finite. Further, there is a function F^* such that for every $\theta \geq 0$ in L^1 , $C^*(\theta) = F^*(\lambda)$ where $\lambda = \int_0^L \theta(x) dx$.*

Proof. The proof of this result involves showing that $C^*(\theta)$ is equal to

$$(3.2a) \quad \min_{u, \phi, y} C(u, y)$$

subject to

$$(3.2b) \quad \begin{aligned} y' &= \bar{A}(y) \bar{u}(x) = \begin{bmatrix} A(y) & b(y) \end{bmatrix} \begin{bmatrix} u(x) \\ \phi(x) \end{bmatrix}, & y(0) &\in K \\ \begin{bmatrix} u(x) \\ \phi(x) \end{bmatrix} &\geq 0, & \int_0^L \begin{bmatrix} u(x) \\ \phi(x) \end{bmatrix} dx &= \begin{bmatrix} a \\ \lambda \end{bmatrix} \geq 0. \end{aligned}$$

Once this is done, the quantity $C^*(\theta)$ depends only on the parameters on which the optimisation problem (3.2) depends; in particular, it does not depend on the form of the function θ , but only on $\lambda = \int_0^L \theta(x) dx$.

First, by Theorem 2.4 the minimum for (3.2) can be obtained at some (u^*, ϕ^*, y^*) . But as any triple (u, θ, y) that satisfies (3.1) also satisfies (3.2), it follows that $C(u^*, y^*) \leq C(u, y)$ and so $C(u^*, y^*) \leq C^*(\theta)$. The infimum therefore is finite.

We now show that $C^*(\theta) = C(u^*, y^*)$. Define $F^*(\lambda)$ to be the optimal value of (3.1) for given $\lambda \geq 0$.

To show this equality holds for $C^*(\theta)$ it suffices to show that there is a sequence (u_ϵ, y_ϵ) as $\epsilon \downarrow 0$ such that $(u_\epsilon, \theta, y_\epsilon)$ satisfies (3.1b) and $C(u_\epsilon, y_\epsilon) \rightarrow C(u^*, y^*)$. This will be done using x -transforms.

Note that if $\lambda = 0$ then $\theta = \phi^* = 0$ almost everywhere, and hence $C(u^*, y^*) = C^*(\theta)$ and there is nothing left to prove. We therefore consider the case that $\lambda > 0$.

Firstly, define

$$\phi_\epsilon^*(x) = \max(\phi^*(x), \epsilon) - \eta(\epsilon)$$

where $\eta(\epsilon)$ is chosen so that

$$\int_0^L \phi_\epsilon^*(x) dx = \int_0^L \theta(x) dx = \lambda.$$

Now $\phi^* > 0$ on a set of positive measure, so $\max(\phi^*, \epsilon) < \phi^* + \epsilon$ on a set of positive measure. Thus

$$\begin{aligned}\lambda &= \int_0^L \phi_\epsilon^*(x) dx = \int_0^L \max(\phi^*(x), \epsilon) dx - \eta(\epsilon)L \\ &< \int_0^L \phi^*(x) dx + (\epsilon - \eta(\epsilon))L \\ &= \lambda + (\epsilon - \eta(\epsilon))L.\end{aligned}$$

Hence $\epsilon > \eta(\epsilon)$ and so $\phi_\epsilon^*(x) \geq \epsilon - \eta(\epsilon) > 0$ for all x . Define

$$\psi_\epsilon(x) = \int_0^x \phi_\epsilon^*(\xi) d\xi, \quad \psi(x) = \int_0^x \theta(\xi) d\xi.$$

Now both ψ and ψ_ϵ are nondecreasing AC functions with

$$\psi'_\epsilon = \phi_\epsilon^* \geq \epsilon - \eta(\epsilon) > 0 \quad \text{a.e.}$$

Thus ψ_ϵ is invertible and ψ_ϵ^{-1} is Lipschitz with constant $1/(\epsilon - \eta(\epsilon))$. Now consider the transformation

$$\tilde{\psi}_\epsilon = \psi_\epsilon^{-1} \circ \psi: [0, L] \rightarrow [0, L].$$

This is well defined as the range of ψ is $[0, \lambda]$, which is the domain of ψ_ϵ^{-1} . Furthermore, ψ_ϵ is AC and nondecreasing as both ψ_ϵ^{-1} and ψ are both nondecreasing and AC [8, Ex. 6, p. 333]. Finally, noting that $\tilde{\psi}_\epsilon(0) = 0$ and $\tilde{\psi}_\epsilon(L) = L$, we see that $\tilde{\psi}_\epsilon$ is also surjective.

We can now set

$$\begin{aligned}u_\epsilon(x) &= u^*(\tilde{\psi}_\epsilon(x)) \cdot \tilde{\psi}'_\epsilon(x) \quad \text{for a.a. } x \in [0, L] \\ \theta_\epsilon(x) &= \phi_\epsilon^*(\tilde{\psi}_\epsilon(x)) \cdot \tilde{\psi}'_\epsilon(x) \quad \text{for a.a. } x \in [0, L] \\ y_\epsilon(x) &= y_\epsilon^*(\tilde{\psi}_\epsilon(x)) \quad \text{for all } x \in [0, L]\end{aligned}$$

where y_ϵ^* is defined through the differential equation

$$y_\epsilon^{*'} = A(y_\epsilon^*) u^*(x) + b(y_\epsilon^*) \phi_\epsilon^*(x), \quad y_\epsilon^*(0) = y^*(0).$$

Thus $(u_\epsilon, \theta_\epsilon, y_\epsilon)$ is an x -transform of $(u^*, \phi_\epsilon^*, y_\epsilon^*)$, and so by Lemma 2.2, $C(u_\epsilon, y_\epsilon) = C(u^*, y_\epsilon^*)$ and $(u_\epsilon, \theta_\epsilon, y_\epsilon)$ satisfies (3.1b).

We now show that $\theta_\epsilon = \theta$ almost everywhere. To do this we apply the chain rule for AC functions [8, Thm. 6.95, p. 325] to $\tilde{\psi}_\epsilon = \psi_\epsilon^{-1} \circ \psi$. This gives $\tilde{\psi}'_\epsilon(x) = \theta(x)/\phi_\epsilon^*(\tilde{\psi}_\epsilon(x))$ for almost all x . Substituting this into the above formula for θ_ϵ then shows that $\theta = \theta_\epsilon$ almost everywhere.

To complete the proof, we now note that $\phi_\epsilon^* \rightarrow \phi^*$ uniformly, and so $y_\epsilon^* \rightarrow y^*$ uniformly, as $\epsilon \downarrow 0$. Hence

$$\lim_{\epsilon \downarrow 0} C(u_\epsilon, y_\epsilon) = \lim_{\epsilon \downarrow 0} C(u_\epsilon^*, y_\epsilon^*) = C(u^*, y^*)$$

as is required. \square

It should also be noted that F^* is, in fact, a continuous function. This follows from the more general result below.

THEOREM 3.2. *In (2.1) the minimum value of $C(u, y)$ is a continuous function of $a \in (\mathbf{R}_+)^m$.*

Proof. We reformulate (2.1) by setting $D = \text{diag}(a_1, \dots, a_m)$ and $u = D\eta$. Then (2.1) becomes the problem of realising

$$(3.3a) \quad \min_{\eta, y} C(\eta, y, a) \quad \text{where } C(\eta, y, a) = g(y(L)) + \int_0^L d(y(x))^T D \eta(x) dx$$

subject to

$$(3.3b) \quad \begin{aligned} y' &= A(y)D\eta, & y(0) &\in K \\ \eta(x) &\geq 0, & \int_0^L \eta(x) dx &= 1 \quad \text{for } i = 1, \dots, m \\ \sum_{i=1}^m \eta_i(x) &= m/L & \text{for all } x &\in [0, L]. \end{aligned}$$

The last equality can be included by Lemma 2.2 and Lemma 2.3.

We now need to show that the minimum value for (3.3) is a continuous function of $a \in (\mathbf{R}_+)^m$. This follows from a standard argument using the compactness of the set of admissible (η, y) pairs and continuity of C .

Firstly, the set of admissible (η, y) is compact, where η is in L^1 with the weak* topology, and y is in $C[0, L]$ with the usual strong topology. Further, $C(\eta, y, a)$ is continuous in $L^1[0, L] \times C[0, L] \times \mathbf{R}^m$ with these topologies. Thus $\min\{C(\eta, y, a) \mid (\eta, y) \text{ satisfy (3.3b)}\}$ depends continuously on a . \square

4. Properties of x -transforms and x -equivalences. Consider again a decomposable system (1.2) where we can compute the $F^*(\cdot)$ function. Suppose further that we know the optimal λ that maximizes $F^*(\lambda)$ and a pair (u_1, y_1) that achieves $\lambda = \int_0^L c(y_1(x))^T u_1(x) dx$. Put $\theta_1 = c(y_1)^T u_1$ and let (u_2, θ_2, y_2) be the optimal solution of the system with input (3.1). If θ_1 and θ_2 are not equal, then the results obtained so far allow us to compute the optimal value for (1.2), but not the controls that achieve it. We need some way of “matching” θ_1 and θ_2 . To do this “matching” step we need to extend the relation of “ x -transforms” to an equivalence relation, which we call “ x -equivalence.”

DEFINITION 4.1. A pair (u, y) is said to be x -equivalent to (v, z) if they have an x -transform in common. That is, there are AC nondecreasing surjective functions $\phi_1, \phi_2: [0, L] \rightarrow [0, L]$ such that

$$\begin{aligned} u(\phi_1(x)) \cdot \phi_1'(x) &= v(\phi_2(x)) \cdot \phi_2'(x) & \text{for a.a. } x \in [0, L] \\ y(\phi_1(x)) &= z(\phi_2(x)) & \text{for all } x \in [0, L]. \end{aligned}$$

Similarly, u and v are said to be x -equivalent if only the former equality holds.

We will show that θ_1 and θ_2 are, in fact, x -equivalent since they are both non-negative and have the same integral on $[0, L]$. Then by applying the appropriate x -transforms to (u_1, y_1) and (u_2, θ_2, y_2) respectively we can obtain a “matched” pair of systems and controls so that the complete system (1.2) can be solved. But to prove the x -equivalence of θ_1 and θ_2 we need the following result.

THEOREM 4.2. *If $\phi, \psi: [0, L] \rightarrow [0, a]$ are two continuous nondecreasing surjective functions then there are nondecreasing surjective Lipschitz functions $\rho, \sigma: [0, L] \rightarrow [0, L]$ such that*

$$\phi \circ \rho = \psi \circ \sigma.$$

Proof. We construct a sequence of pairs of functions $\rho_n, \sigma_n: [0, L] \rightarrow [0, L]$ that are Lipschitz with constant 2. Let $h_n = L/2^n$. We now construct $\rho_n(k h_n)$ and $\sigma_n(k h_n)$ for $k = 0, 1, \dots, 2^n$ by induction, and the value of ρ_n and σ_n on the remainder of $[0, L]$ is defined by linear interpolation on these values. We will prove the following, for our construction, by induction on k :

- (1) $\phi(\rho_n(k h_n)) = \psi(\sigma_n(k h_n))$.
- (2) $|\rho_n((k+1)h_n) - \rho_n(k h_n)| \leq 2 h_n$ and $|\sigma_n((k+1)h_n) - \sigma_n(k h_n)| \leq 2 h_n$.
- (3) $\rho_n((k+1)h_n) \geq \rho_n(k h_n)$ and $\sigma_n((k+1)h_n) \geq \sigma_n(k h_n)$.

For $k = 0$ we set $\rho_n(0) = \sigma_n(0) = 0$. Then (1) holds for $k = 0$ as $\phi(0) = \psi(0) = 0$.

Now suppose that $\rho_n(k h_n)$ and $\sigma_n(k h_n)$ have been constructed. We construct $\rho_n((k+1)h_n)$ and $\sigma_n((k+1)h_n)$ according to the following cases:

Case 1. If $\rho_n(k h_n) + 2h_n, \sigma_n(k h_n) + 2h_n \geq L$ then put $\rho_n((k+1)h_n) = \sigma_n((k+1)h_n) = L$. Then (1) holds as $\phi(L) = \psi(L) = a$. Items (2) and (3) hold by inspection.

Case 2a. If $\rho_n(k h_n) + 2h_n < L$ but $\sigma_n(k h_n) + 2h_n \geq L$ then put

$$\rho_n((k+1)h_n) = \rho_n(k h_n) + 2h_n$$

$$\sigma_n((k+1)h_n) = \min\{s \mid \psi(s) = \phi(\rho_n((k+1)h_n)) \text{ and } s \geq \sigma_n(k h_n)\}.$$

As $\psi(L) = \phi(L) \geq \phi(\rho_n((k+1)h_n))$ and $\psi(\sigma_n(k h_n)) = \phi(\rho_n(k h_n)) \leq \phi(\rho_n((k+1)h_n))$, a minimising $s \leq L$ must exist, and (2) holds. Item (1) follows directly from the definition of $\sigma_n((k+1)h_n)$ and (3) is apparent on inspection.

Case 2b. If $\sigma_n(k h_n) + 2h_n < L$ but $\rho_n(k h_n) + 2h_n \geq L$ then apply Case 2a, but with the roles of ρ_n, σ_n and ϕ, ψ reversed.

Case 3. Assume that $\rho_n(k h_n) + 2h_n, \sigma_n(k h_n) + 2h_n < L$, and assume without loss of generality that

$$\phi(\rho_n(k h_n) + 2h_n) \leq \psi(\sigma_n(k h_n) + 2h_n).$$

(Otherwise reverse the roles of ρ_n, σ_n and ϕ, ψ .) Then put

$$\rho_n((k+1)h_n) = \rho_n(k h_n) + 2h_n$$

$$\sigma_n((k+1)h_n) = \min\{s \mid \psi(s) = \phi(\rho_n((k+1)h_n)) \text{ and } s \geq \sigma_n(k h_n)\}.$$

By the same reasoning as in Case 2 there must be a minimising s where $\sigma_n(k h_n) \leq s \leq \sigma_n(k h_n) + 2h_n$, so (2) holds. Items (1) and (3) hold as in Case 2.

We now show that $\rho_n(k h_n) + \sigma_n(k h_n) \geq 2h_n$ by induction. Note that from this it follows that $\rho_n(L) + \sigma_n(L) = 2L$, and so $\rho_n(L) = \sigma_n(L) = L$.

For $k = 0$ the result is true since $\rho_n(0) = \sigma_n(0) = 0$.

Suppose true for k ; we now show the assertion holds for $k+1$. In Case 1, $\rho_n((k+1)h_n) + \sigma_n((k+1)h_n) = L + L \geq (k+1)h_n$. In Cases 2 and 3 either

$$\rho_n((k+1)h_n) = \rho_n(k h_n) + 2h_n \quad \text{or} \quad \sigma_n((k+1)h_n) = \sigma_n(k h_n) + 2h_n.$$

Since both ρ_n and σ_n are nondecreasing, the desired result follows.

Noting that $\rho_n, \sigma_n: [0, L] \rightarrow [0, L]$ are nondecreasing, surjective functions with Lipschitz constant 2, by the Arzelà–Ascoli theorem there must be a convergent subsequence with a limit (ρ, σ) . By the properties of uniform convergence, both ρ and σ are nondecreasing, surjective functions $[0, L] \rightarrow [0, L]$.

To show that $\phi \circ \rho = \psi \circ \sigma$, let $t = kL/2^n$ for some $k = 0, \dots, 2^n$. Then for $m \geq n$, $\phi(\rho_m(t)) = \psi(\sigma_m(t))$ and taking $m \rightarrow \infty$ in the above subsequence gives

$$\phi(\rho(t)) = \psi(\sigma(t)).$$

By continuity of ϕ, ψ, ρ and σ and that $\{kL/2^n \mid k = 0, \dots, 2^n, n \geq 0\}$ is dense in $[0, L]$ it follows that $\phi \circ \rho = \psi \circ \sigma$ everywhere. \square

COROLLARY 4.3. If $\theta, \eta: [0, L] \rightarrow \mathbf{R}_+$ are in $L^1[0, L]$ and

$$\int_0^L \theta(x) dx = \int_0^L \eta(x) dx$$

then θ and η are x -equivalent.

Proof. Let

$$\phi(x) = \int_0^x \theta(\xi) d\xi, \quad \psi(x) = \int_0^x \eta(\xi) d\xi.$$

Then by Theorem 4.2, there are nondecreasing, surjective Lipschitz functions $\rho, \sigma: [0, L] \rightarrow [0, L]$ such that

$$\phi \circ \rho = \psi \circ \sigma.$$

These composite functions are AC, since ρ and σ are both AC and nondecreasing. Differentiating this equation by the chain rule gives

$$\phi'(\rho(x)) \cdot \rho'(x) = \psi'(\sigma(x)) \cdot \sigma'(x) \quad \text{for a.a. } x$$

which is the desired x -equivalence. \square

Thus, θ_1 and θ_2 are x -equivalent as we wished to show. The following result is of independent interest.

COROLLARY 4.4. The relation “ x -equivalence” is an equivalence relation.

Proof. From the definition of “ x -equivalence” it is clear that it is both reflexive ((u, y) is always x -equivalent to (u, y)) and symmetric ((u, y) is x -equivalent to (v, z) implies the reverse x -equivalence). To complete the proof it suffices to show that “ x -equivalence” is a transitive relation.

Let (u_1, y_1) and (u_2, y_2) be x -equivalent; (u_2, y_2) and (u_3, y_3) be x -equivalent. We show that (u_1, y_1) and (u_3, y_3) are x -equivalent. By definition of x -equivalence there exist AC nondecreasing, surjective functions $\rho_1, \rho_2, \sigma_1, \sigma_2: [0, L] \rightarrow [0, L]$ such that

$$\begin{aligned} y_1(\rho_1(x)) &= y_2(\sigma_1(x)) & y_2(\rho_2(x)) &= y_3(\sigma_2(x)) \\ u_1(\rho_1(x))\rho_1'(x) &= u_2(\sigma_1(x))\sigma_1'(x) & u_2(\rho_2(x))\rho_2'(x) &= u_3(\sigma_2(x))\sigma_2'(x) \end{aligned}$$

for almost all $x \in [0, L]$. By Theorem 4.2 there are AC nondecreasing surjective functions $\omega_1, \omega_2: [0, L] \rightarrow [0, L]$ where $\sigma_1 \circ \omega_1 = \rho_2 \circ \omega_2$. Then if $\phi_1 = \rho_1 \circ \omega_1$ and $\phi_2 = \sigma_2 \circ \omega_2$ we find that

$$\begin{aligned} y_1(\phi_1(x)) &= y_1(\rho_1(\omega_1(x))) \\ &= y_2(\sigma_1(\omega_1(x))) = y_2(\rho_2(\omega_2(x))) \\ &= y_3(\sigma_2(\omega_2(x))) = y_3(\phi_2(x)). \end{aligned}$$

Similar arguments will show that $u_1(\phi_1(x))\phi_1'(x) = u_3(\phi_2(x))\phi_2'(x)$ by means of the chain rule for AC functions. \square

We can now return to the problem of how analysis of the two subsystems can be combined to provide a complete solution for the entire system (1.2). Note first, that whatever u_1 and y_1 are chosen satisfying (1.2b), the infimum of $C(u_2, y_2)$ over (u_2, y_2) satisfying (1.2b) depends only on $\lambda = \int_0^L \theta(x) dx = \int_0^L c(y_1(x))^T u_1(x) dx$. What, then, is the set of attainable values of λ ? By continuous dependence arguments it is easy to see that λ is a continuous function of (u_1, y_1) , weakly in u_1 . Further, by the results of §2 the set of admissible (u_1, y_1) is compact in $L^1[0, L] \times C[0, L]$, with $L^1[0, L]$ having the weak* topology. Also, as K_1 is connected, so is the set of admissible pairs (u_1, y_1) . Thus the set of attainable values of λ is a compact connected set; that is, this

set is a finite closed interval $[\lambda_{\min}, \lambda_{\max}]$. The extreme values λ_{\min} and λ_{\max} can be determined by solving the appropriate optimisation problems for subsystem S_1 :

$$\max / \min \int_0^L c(y_1(x))^T u_1(x) dx$$

subject to

$$\begin{aligned} y_1' &= A_1(y_1) u_1, & y_1(0) &\in K_1 \\ u_1(x) &\geq 0, & \int_0^L u_1(x) dx &= a_1. \end{aligned}$$

These problems are of the form of (2.1), and so by (3.2) solutions exist.

Now, for any value of λ , the infimum of $C(u_2, y_2)$ over (u_2, y_2) satisfying (1.2b) is given by $F^*(\lambda)$. We have shown that F^* is a continuous function, and as λ ranges over a compact set $[\lambda_{\min}, \lambda_{\max}]$ this function must have a minimum value $F^*(\lambda^*)$. By using the pathwise connectedness of the set of admissible (u_1, y_1) we can construct a pair (\bar{u}_1, \bar{y}_1) that satisfies (1.2b) and $\int_0^L c(\bar{y}_1(x))^T \bar{u}_1(x) dx = \lambda^*$. We also have a triple $(\bar{u}_2, \theta_2, \bar{y}_2)$ that solves (3.1) with $\lambda = \lambda^*$.

As

$$\int_0^L \theta_1(x) dx = \int_0^L \theta_2(x) dx = \lambda^*$$

θ_1 and θ_2 are x -equivalent; let $\phi_1, \phi_2: [0, L] \rightarrow [0, L]$ be two AC nondecreasing, surjective functions such that

$$(\theta_1 \circ \phi_1) \cdot \phi_1' = (\theta_2 \circ \phi_2) \cdot \phi_2' \quad \text{a.e.}$$

Then setting

$$\begin{aligned} y_1^*(x) &= \bar{y}_1(\phi_1(x)) & y_2^*(x) &= \bar{y}_2(\phi_2(x)) \\ u_1^*(x) &= \bar{u}_1(\phi_1(x))\phi_1'(x) & u_2^*(x) &= \bar{u}_2(\phi_2(x))\phi_2'(x) \end{aligned}$$

gives a solution to (1.2); $(u_1^*, u_2^*, y_1^*, y_2^*)$ satisfies (1.2b) and $C(u_2^*, y_2^*) = F^*(\lambda^*)$ is the required minimum.

5. Results for one-enzyme systems. Under the assumptions described in the introduction for a liver tubule or a packed-bed reactor, where diffusion along the flow is negligible, a one-enzyme subsystem is described by a pair of *ODEs*, one describing the concentration of the reactant R , and another for the product P , which are assumed to have reached steady state. The enzyme concentration at a distance x from the entrance to the reactor is $f(x)$; the reaction rate per unit enzyme is given by $\alpha(R)$; and we allow an "input" (possibly zero) of reactant denoted $\theta(x)$ per unit length. The *ODEs* are then

$$\begin{aligned} (5.1) \quad dR/dx &= \theta(x) - f(x)\alpha(R) \\ dP/dx &= f(x)\alpha(R). \end{aligned}$$

It is assumed that α satisfies the following properties:

$$\begin{aligned} \alpha(0) &= 0 \\ R > 0 &\implies \alpha(R) > 0 \\ \alpha &\text{ is a Lipschitz } C^1 \text{ function.} \end{aligned}$$

Also, f is assumed to satisfy

$$f(x) \geq 0 \quad \text{for all } x \in [0, L]$$

$$f \in L^1[0, L], \quad \int_0^L f(x) dx = q \geq 0.$$

This “input” θ may be the product of some other reaction or reactions.

Here we consider the problem of realising

$$(5.2a) \quad \max_{f, \theta} P(L)$$

subject to (5.1) and

$$(5.2b) \quad \begin{aligned} R(0) &= R_0, & P(0) &= P_0 \\ f(x) &\geq 0, & \theta(x) &\geq 0 \quad \text{for all } x \in [0, L], \\ \int_0^L f(x) dx &= q, & \int_0^L \theta(x) dx &= \lambda. \end{aligned}$$

Since $P(L) + R(L) = P(0) + R(0) + \lambda$, maximising $P(L)$ is equivalent to minimising $R(L)$.

The work of this section uses the ideas and analysis of Holm aker and Stewart[6] rather heavily, although [6] did not deal with problems where $R_0 > 0$. Consequently, some additional results are proven to complete the application.

Let

$$G^*(\lambda, q) = \min_{f, \theta} R(L) \quad \text{subject to (5.2b)}$$

$$F^*(\lambda, q) = \max_{f, \theta} P(L) \quad \text{subject to (5.2b)}.$$

In this section we show that

$$F^*(\lambda, q) = \max_{0 \leq \rho \leq \lambda + R_0} F(\rho, \lambda, q)$$

$$G^*(\lambda, q) = \min_{0 \leq \rho \leq \lambda + R_0} G(\rho, \lambda, q)$$

$$\text{where } F(\rho, \lambda, q) = \lambda + P_0 + R_0 - G(\rho, \lambda, q)$$

and $G(\rho, \lambda, q)$ is given by Table 5.1.

TABLE 5.1

Case	Conditions		$G(\rho, \lambda, q)$
1	$R_0 \leq \rho$	$\rho \geq \lambda + R_0$	$\int_G^{R_0 + \lambda} dr/\alpha(r) = q$
2	$R_0 \leq \rho$	$\rho + q\alpha(\rho) \geq \lambda + R_0 \geq \rho$	$\int_G^\rho dr/\alpha(r) = q - (\lambda + R_0 - \rho)/\alpha(\rho)$
3	$R_0 \leq \rho$	$\lambda + R_0 \geq \rho + q\alpha(\rho)$	$\lambda + R_0 - q\alpha(\rho)$
4	$R_0 \geq \rho$	$q \geq \lambda/\alpha(\rho) + \int_\rho^{R_0} dr/\alpha(r)$	$\int_G^{R_0} dr/\alpha(r) = q - \lambda/\alpha(\rho)$
5	$R_0 \geq \rho$	$\lambda/\alpha(\rho) + \int_\rho^{R_0} dr/\alpha(r) \geq q \geq \int_\rho^{R_0} dr/\alpha(r)$	$\lambda + \rho - \alpha(\rho) \left[q - \int_\rho^{R_0} dr/\alpha(r) \right]$
6	$R_0 \geq \rho$	$\rho \leq \rho_{\min}$	$\lambda + \rho_{\min}$

Here ρ_{\min} is defined by

$$\int_{\rho_{\min}}^{R_0} \frac{dr}{\alpha(r)} = q.$$

By inspection it is clear that $G(\rho, \lambda, q)$ is well defined for $\rho, \lambda, q \geq 0$.

First it should be noted that the values given for $G(\rho, \lambda, q)$ are attainable for each $\rho \in \mathbf{R}_+$. To demonstrate this, the following list that shows how to construct θ, f given (ρ, λ, q) so that the corresponding solution $R_\rho(x)$ satisfies $R_\rho(L) = G(\rho, \lambda, q)$. For the construction to proceed we need $0 < \xi_1 < \xi_2 < L$ and the intervals $I_1 = [0, \xi_1]$, $I_2 = [\xi_1, \xi_2]$ and $I_3 = [\xi_2, L]$.

Case 1. Put $\theta \geq 0$ and $f = 0$ on I_1 , and $\theta = 0$ and $f \geq 0$ on $I_2 \cup I_3$ so that $\int_{I_1} \theta = \lambda$ and $\int_{I_2 \cup I_3} f = q$.

Case 2. Put $\theta \geq 0$ and $f = 0$ on I_1 , with $\theta = \alpha(\rho)f \geq 0$ on I_2 and $\theta = 0$, $f \geq 0$ on I_3 so that $\int_{I_1} \theta = \rho$, $\int_{I_2} f = (\lambda + R_0 - \rho)/\alpha(\rho) \leq q$ and (consequently) $\int_{I_3} f = q - (\lambda + R_0 - \rho)/\alpha(\rho)$.

Case 3. Put $\theta \geq 0$ and $f = 0$ on $I_1 \cup I_3$ and $\theta = \alpha(\rho)f \geq 0$ on I_2 so that $\int_{I_1} \theta = \rho$, $\int_{I_2} f = q$ and $\int_{I_3} \theta = \lambda - \rho$.

Case 4. Put $\theta = 0$ and $f \geq 0$ on $I_1 \cup I_3$, $\theta = \alpha(\rho)f \geq 0$ on I_2 so that $\int_{I_1} f = \int_{\rho}^{R_0} dr/\alpha(r)$, $\int_{I_2} \theta = \lambda$ and $\int_{I_3} f = q - \lambda/\alpha(\rho) - \int_{\rho}^{R_0} dr/\alpha(r)$.

Case 5. Put $\theta = 0$ and $f \geq 0$ on I_1 , $\theta = \alpha(\rho)f \geq 0$ on I_2 , and $\theta \geq 0$ and $f = 0$ on I_3 so that $\int_{I_1} f = \int_{\rho}^{R_0} dr/\alpha(r)$, $\int_{I_2} f = q - \int_{\rho}^{R_0} dr/\alpha(r)$ and $\int_{I_3} \theta = \lambda - \alpha(\rho)(q - \int_{\rho}^{R_0} dr/\alpha(r))$.

Case 6. Put $\theta = 0$ and $f \geq 0$ on I_1 and $\theta \geq 0$, $f = 0$ on $I_2 \cup I_3$ so that $\int_{I_1} f = q$ and $\int_{I_2 \cup I_3} \theta = \lambda$.

That $R_\rho(L) = G(\rho, \lambda, q)$, and that admissible f and θ can be constructed according to the above rules, can be shown by inspection. As noted above, Cases 1–3 are proven in Holm aker and Stewart [7].

The problem now is to show that *any* solution $R(x)$ of (5.2.b) for admissible f and θ satisfies $R(L) \geq R_\rho(L)$ for some $\rho \in [0, \lambda]$. This is the substance of the following theorem.

THEOREM 5.1. *Let $G(\rho, \lambda, q)$ be defined as in Table 5.1 for $\rho, \lambda, q \geq 0$ and $F(\rho, \lambda, q) = \lambda + P_0 + R_0 - G(\rho, \lambda, q)$. Then the optimal values for $P(L)$ and $R(L)$ are given by*

$$\begin{aligned} F^*(\lambda, q) &= \max_{0 \leq \rho \leq \lambda + R_0} F(\rho, \lambda, q), \\ G^*(\lambda, q) &= \min_{0 \leq \rho \leq \lambda + R_0} G(\rho, \lambda, q), \quad \text{respectively.} \end{aligned}$$

Proof. We have already noted that $P(L) + R(L) = P_0 + R_0 + \lambda$, so that maximising $P(L)$ corresponds to minimising $R(L)$. Now $\min_{f, \theta} R(L) \leq G^*(\lambda, q)$ since we can construct a solution to (5.2.b) $R_{\rho^*}(x)$ that achieves $R_{\rho^*}(L) = G(\rho^*, \lambda, q) = G^*(\lambda, q)$.

We now show that the reverse inequality for $\min_{f, \theta} R(L)$ holds. Let $\alpha^* = \max\{\alpha(R(x)) \mid x \in [0, L]\}$ and $\xi = \max\{x \mid \alpha(R(x)) = \alpha^*\}$. Put $\rho = R(\xi)$. First, we note that if $\lambda > 0$ then $\rho > 0$, as otherwise $R(x) = 0$ for all x , so $\theta = 0$ almost everywhere, and $\lambda = 0$. In the case $\lambda = 0$ it is clear that $R(L)$ is independent of the particular admissible control f chosen.

The proofs for Cases 1–3 now follows Holm aker and Stewart [7, Thm. 5.1], with λ replaced by $\lambda + R_0$. We now consider in detail Cases 4–6; that is, $R_0 \geq \rho$ and so $R(x) > 0$ for all x .

Case 4. Here $R_\rho(L)$ is given by

$$\int_{R_\rho(L)}^{R_0} \frac{dr}{\alpha(r)} = q - \frac{\lambda}{\alpha(\rho)}.$$

However, for R ,

$$(5.3) \quad \frac{1}{\alpha(R(x))} \frac{dR}{dx} = \frac{\theta(x)}{\alpha(R(x))} - f(x).$$

Integrating this last equation from 0 to L gives

$$\begin{aligned} - \int_{R(L)}^{R_0} \frac{dr}{\alpha(r)} &= \int_0^L \frac{\theta(x) dx}{\alpha(R(x))} - \int_0^L f(x) dx \\ &\geq \frac{\lambda}{\alpha(\rho)} - q = - \int_{R_\rho(L)}^{R_0} \frac{dr}{\alpha(r)} \end{aligned}$$

so that $R_\rho(L) \leq R(L)$.

Case 5. Integrating (5.3) from 0 to ξ gives

$$(5.4) \quad \begin{aligned} - \int_\rho^{R_0} \frac{dr}{\alpha(r)} &= \int_0^\xi \frac{\theta(x) dx}{\alpha(R(x))} - \int_0^L f(x) dx \\ &\geq \frac{1}{\alpha(\rho)} \int_0^\xi \theta(x) dx - \int_0^L f(x) dx. \end{aligned}$$

Let $q' = q - \int_\rho^{R_0} dr/\alpha(r)$. Then integrating the ODE for R in (5.2.b) from ξ to L we find that

$$\begin{aligned} R(L) - \rho &= \int_\xi^L \theta(x) dx - \int_\xi^L \alpha(R(x)) f(x) dx \\ &\geq \int_\xi^L \theta(x) dx - \alpha(\rho) \int_\xi^L f(x) dx \\ &\geq \int_0^L \theta(x) dx - \alpha(\rho) q' \quad \text{by (5.4)} \\ &= \lambda - \alpha(\rho) q' \end{aligned}$$

so that $R(L) \geq \lambda + \rho - \alpha(\rho) q' = R_\rho(L)$ as required.

Case 6. Noting that integrating (5.3) from 0 to ξ gives the result that

$$(5.5) \quad \int_\rho^{R_0} \frac{dr}{\alpha(r)} \leq q.$$

Thus Case 6 can only occur if (5.5) is an equality, so that Case 5 may be applied. Noting that for $\rho < \rho_{\min}$ implies that $G(\rho, \lambda, q) = G(\rho_{\min}, \lambda, q)$, we see that this possibility does not affect the minimisation. \square

A stronger characterisation of the optimising value of ρ for $F(\rho, \lambda, q)$ can be obtained; this will allow the optimal value ρ^* to be computed almost directly.

THEOREM 5.2. *There is a minimizer $\rho^* \in [\rho_{\min}, \lambda + R_0]$ of $G(\cdot, \lambda, q)$ such that*

$$\alpha(\rho^*) = \max\{\alpha(\rho) \mid \rho \text{ lies between } R_0 \text{ and } \rho^*\}.$$

If, in addition, $\alpha \in C^1$ and $\rho^ \neq \lambda + R_0$, ρ_{\min} , then $\alpha'(\rho^*) = 0$.*

Proof. We prove the first part of the theorem. Let $G(\hat{\rho}, \lambda, q) = G^*(\lambda, q)$. Let $R_{\hat{\rho}}$ be defined as a solution of (5.2.b) with f and θ constructed as described above. Let $\rho^* = R_{\hat{\rho}}(\xi)$ where

$$\alpha(R_{\hat{\rho}}(\xi)) = \max\{\alpha(R_{\hat{\rho}}(x)) \mid x \in [0, L]\}.$$

Note that $R_{\hat{\rho}}(x) \in [\rho_{\min}, \lambda + R_0]$ for all x , so ρ^* lies in the required interval.

By the arguments used in Theorem 5.1 with ρ replaced by ρ^* and R replaced by $R_{\hat{\rho}}$, we find that $R_{\rho^*}(L) \leq R_{\hat{\rho}}(L)$. However, as $\hat{\rho}$ minimizes $G(\cdot, \lambda, q)$, it follows that $R_{\rho^*}(L) = R_{\hat{\rho}}(L)$ and ρ^* is another minimizer of G . Now $R_{\hat{\rho}}$ attains both R_0 and ρ^* , so that by definition of ρ^*

$$\alpha(\rho^*) \geq \max\{\alpha(\rho) \mid \rho \text{ lies between } R_0 \text{ and } \rho^*\}.$$

Setting $\rho = \rho^*$ in the maximisation makes it clear that the above is actually an equality.

We now show the second part of the theorem. Suppose that $\alpha \in C^1$ and $\rho^* \neq \lambda + R_0, \rho_{\min}$; we now show that $\alpha'(\rho^*) = 0$. To do this we note that the derivatives of $G(\rho, \lambda, q)$ with respect to ρ are as in Table 5.2.

TABLE 5.2

Case	$\partial G / \partial \rho(\rho, \lambda, q)$
1	0
2	$-(\lambda + R_0 - \rho)\alpha(G)\alpha'(\rho)/\alpha(\rho)^2$
3	$-q\alpha'(\rho)$
4	$-\lambda\alpha(G)\alpha'(\rho)/\alpha(\rho)^2$
5	$-\alpha'(\rho)\left[q - \int_{\rho}^{R_0} dr/\alpha(r)\right]$
6	0

By inspection of Table 5.2 and 5.1 $\partial G / \partial \rho$ is continuous except at $\rho = \lambda + R_0, \rho_{\min}$. Thus, as $\rho = \rho^*$ minimizes $G(\rho, \lambda, q)$ it follows that $\partial G / \partial \rho(\rho^*, \lambda, q) = 0$. Hence, by Table 5.2, $\alpha'(\rho^*) = 0$ as required. \square

The above formulae in 5.1 provide a fairly easy way of computing $G^*(\lambda, q)$ and $F^*(\lambda, q)$. However, F^* and G^* have some monotonicity properties that will be exploited later.

THEOREM 5.3. *Both $F^*(\lambda, q)$ and $G^*(\lambda, q)$ are monotonic increasing in λ with G^* being strictly increasing in λ . Furthermore, $F^*(\lambda, q)$ is strictly increasing in q and $G^*(\lambda, q)$ is strictly decreasing in q .*

Proof. Let $0 \leq \lambda_1 < \lambda_2$ and suppose that (θ, f) achieves $P(L) = F^*(\lambda_1, q)$ where P and R satisfy (5.1). We then define

$$\bar{\theta}(x) = \begin{cases} 2\theta(2x) & \text{for } 0 \leq x \leq L/2 \\ 2(\lambda_2 - \lambda_1)/L & \text{otherwise,} \end{cases}$$

$$\bar{f}(x) = \begin{cases} 2f(2x) & \text{for } 0 \leq x \leq L/2 \\ 0 & \text{otherwise,} \end{cases}$$

and let \bar{P} and \bar{R} satisfy (5.1) with (θ, f) replaced by $(\bar{\theta}, \bar{f})$. Then $\bar{P}(L) = P(L) = F^*(\lambda_1, q)$, so the maximum value of $P(L)$ over (θ, f) with $\int_0^L \theta(x) dx = \lambda_2$ must be at least as large as $F^*(\lambda_1, q)$; that is, $F^*(\lambda_2, q) \geq F^*(\lambda_1, q)$ as required.

On the other hand, suppose now that (θ, f) achieves $R(L) = G^*(\lambda_2, q)$. Then there is a $\xi \in (0, L)$ such that $\int_0^{\xi} \theta(x) dx = \lambda_1$. If $R_0 = 0$ and $\lambda_1 = 0$ the result follows trivially as then $R(L) = 0$ and thus $G^*(\lambda_1, q) = 0 < G^*(\lambda_2, q)$. We now assume that either R_0 or λ_1 is strictly positive; in this case $R(x) > 0$ for all $x \in [\xi, L]$. We define

$$\bar{\theta}(x) = \begin{cases} \theta(x) & \text{for } 0 \leq x \leq \xi \\ 0 & \text{otherwise.} \end{cases}$$

If \bar{P} and \bar{R} then satisfy (5.1) with θ replaced by $\bar{\theta}$ we find that

$$\begin{aligned}\int_{R(L)}^{R(\xi)} \frac{dr}{\alpha(r)} &= \int_{\xi}^L \frac{\theta(x) dx}{\alpha(R(x))} - \int_{\xi}^L f(x) dx \\ &> - \int_{\xi}^L f(x) dx = \int_{R(L)}^{\bar{R}(\xi)} \frac{dr}{\alpha(r)}\end{aligned}$$

so that $\bar{R}(L) < R(L) = G^*(\lambda_2, q)$. Thus $G^*(\lambda_1, q) < G^*(\lambda_2, q)$ as required.

We now consider the effect of varying q ; let $0 \leq q_1 < q_2$ and suppose (θ, f) achieves $P(L) = F^*(\lambda, q_1)$ and $R(L) = G^*(\lambda, q_1)$. We then define

$$\begin{aligned}\bar{\theta}(x) &= \begin{cases} 2\theta(2x) & \text{for } 0 \leq x \leq L/2 \\ 0 & \text{otherwise,} \end{cases} \\ \bar{f}(x) &= \begin{cases} 2f(2x) & \text{for } 0 \leq x \leq L/2 \\ 2(q_2 - q_1)/L & \text{otherwise.} \end{cases}\end{aligned}$$

Then clearly $\bar{P}(L) > P(L) = F^*(\lambda, q_1)$ and $\bar{R}(L) < G^*(\lambda, q_1)$ so that $F^*(\lambda, q_2) > F^*(\lambda, q_1)$ and $G^*(\lambda, q_2) < G^*(\lambda, q_1)$ as required. \square

These monotonicity results considerably simplify the application of the results of §§2–4 to networks of enzymatic reactions.

6. Solving complete networks. In this section we combine the results of §§4 and 5 to obtain a method for maximising the output of a large tree of enzymatic reactions by a dynamic programming approach. (Note that this includes problems where there is a *chain* of reactions, or independent reaction paths.) This method has been implemented numerically to obtain optimal values, and extended to also compute optimal *amounts* of each enzyme subject to a “total cost” limit. Currently the implementation provides only qualitative information about the optimal controls (enzyme distributions) themselves, though this alone is often sufficient to determine optimal distributions (e.g., “... the first enzyme is followed by the second enzyme which is followed by the thirrd ...”).

We consider firstly the problem of maximising output where the quantity of each enzyme is given and the set of reactions forms a tree directed towards a single product: the “output.” If the product is the product of more than one reaction, then the tree can be split into different parts—one for each reaction producing the product—which are entirely independent. Thus we can assume that there is only one reaction producing the final product. This means that the tree of reactions can be split into the form

$$T = T' \Rightarrow R \rightarrow P$$

where “ \Rightarrow ” means that R may be the product of more than one reaction. Now the first subsystem S_1 is the subtree $T' \Rightarrow R$, and the second, S_2 , is the one-enzyme reaction with input: $\theta \rightarrow R \rightarrow P$. (Here θ is the sum of the production rates of R over all reactions that produce R .) The maximum value of $P(L)$ over all θ, f satisfying the nonnegativity and the integral constraints

$$\int_0^L \theta(x) dx = \lambda, \quad \int_0^L f(x) dx = q$$

can be computed as $F^*(\lambda, q)$ by means of the results of the previous section. Noting that $F^*(\lambda, q)$ is monotonic increasing in λ , we see that if the set of attainable λ is

$[\lambda_{\min}, \lambda_{\max}]$, then the optimum value of λ is $\lambda^* = \lambda_{\max}$. But λ_{\max} in this case is simply the maximum output of subsystem S_1 :

$$T' \Rightarrow R.$$

This subsystem can then be analysed in terms of its own subsystems, which in turn can be analysed in the same manner, at least until we are dealing with single enzyme systems with no input, which are essentially trivial: any enzyme distribution satisfying the nonnegativity and integral conditions will achieve the maximum.

The optimal enzyme distributions themselves can be generated with the above recursive process; at each stage, once the optimal λ^* and associated output from the S_1 subsystem θ_1 has been determined, then the optimal input θ_2 can be determined by the results for one enzyme systems. Then the optimal distributions and trajectories for the two subsystems can be “matched” by means of the x -equivalence described at the end of §4.

We now briefly consider a particular case of a chain of enzymes where the reaction rate functions, the α 's, are C^1 and $\alpha'(\rho) > 0$ for every $\rho > 0$. Here we use the notation of the previous section and consider a one enzyme subsystem. By Theorem 5.2 we can pick an optimum ρ^* such that

$$\alpha(\rho^*) = \max\{\alpha(\rho) \mid \rho \text{ lies between } R_0 \text{ and } \rho^*\}.$$

Thus $\rho^* \geq R_0$. As $\alpha'(\rho^*) \neq 0$, we see that $\rho^* = \lambda$ or $\rho^* = \rho_{\min}$. But $\rho_{\min} < R_0$ for $R_0 > 0$, and $\rho_{\min} = 0$ if $R_0 = 0$, it follows that $\rho^* = \lambda$ and we can apply Case 1. We then subdivide the interval $[0, L]$ into two pieces $I_1 \cup I_2 = [0, \xi] \cup [\xi, L]$ and set $f = 0$ on I_1 and $\theta = 0$ on I_2 . That is, we put all of the current enzyme downstream of the enzymes that come earlier in the chain of reactions. Thus, in this case, the naive approach of putting the enzymes in order is actually optimal. For the more general case, we need to resort to numerical methods to determine the optimal distribution of enzymes, which are described below.

This technique can also be extended to give a dynamic programming method for obtaining the optimum *amount* of each enzyme as well as the optimum distribution of a given amount of each enzyme, subject to a limit on a “total enzyme cost” given as a weighted sum of the amounts of the various enzymes. For each reaction subsystem (i.e., subtree) a table of amounts of product for a given total enzyme cost for the subsystem is computed. To do this we have basically two cases to consider. The first is where a product P is produced from a reactant R , which in turn is the product of some system of reactions. The problem here is to decide how to split the total enzyme cost between the reaction $R \rightarrow P$ and the system of reactions that produces R . This is just a one-dimensional optimisation problem.

The second case is where the product P is the product of two independent systems of reactions. Then the problem to be solved is that of determining how to split this total enzyme cost between the two subsystems of reactions. This, again, is a one-dimensional optimisation problem that can be solved directly from the tables of product amounts for the two subsystems.

7. Numerical implementation and results. This decomposition algorithm has been implemented numerically in the C programming language to run on the Mathematics Dept. Pyramid 9810 computer. The method used to compute $G(\rho, \lambda, q)$ involves a one-dimensional root-finding routine based on that of Brent [3, pp. 58–59], and the integrals were computed by a 24-point Gauss–Legendre quadrature rule. Finally, the reaction rate functions α were stored in a form so that their critical points

could be computed directly. Then $F^*(\lambda, q)$ is given directly as the maximum of the $F(\rho, \lambda, q)$ as ρ ranges over the set of critical points of α in the interval $[\rho_{\min}, \lambda + R_0]$ together with $\rho = \lambda + R_0$ and $\rho = \rho_{\min}$.

The optimal amounts of product were computed for fixed amounts of enzymes, and for variable amounts of enzymes, for a problem devised by the author to test the algorithm, which is given in Fig. 7.1.

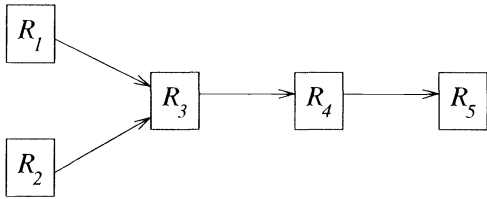


FIG. 1

The reaction rate functions are given in Table 7.1.

TABLE 7.1

Edge	Reaction rate	Amount	Critical points	$R_i(0)$
1	$\alpha_1(r) = 8r/(r^2 + 4)$	2	not used	3
2	$\alpha_2(r) = (r^2 + r)/(r^2 + r + 2)$	2	not used	3
3	$\alpha_3(r) = 6re^{-r}$	2	$\alpha_3(1) = 2.207 \dots$	0
4	$\alpha_4(r) = 4(e^{-r/5} - e^{-2r/5})$	3	$\alpha_4(3.465 \dots) = 1$	0

Note that edge i is the outgoing edge of node R_i .

For determining the optimal amount of each enzyme it was assumed that the total amount of enzyme was the same as for the case where the amount of each enzyme is given. The smallest change in the amount of an enzyme that was allowed by the numerical implementation was set to 0.2.

For making meaningful statements about the efficacy of the algorithm and the optimal amounts and distributions of enzymes, comparisons were made with two heuristics:

- (1) Uniform enzyme concentration.
- (2) Enzymes occur in the sequence determined by the reaction sequence.

The results are given in Table 7.2. The optimal enzyme amounts computed were

TABLE 7.2

Edge	1	2	3	4
Amount	1.6	1.4	1.8	4.2

The amounts of product produced under the different conditions were computed to be as in Table 7.3.

TABLE 7.3

Amounts	Heuristic 1	Heuristic 2	Optimal distribution
Given	1.796	2.283	2.802
Optimal	1.955	2.540	3.166

There is clearly significant advantage in using the computed optimal distributions and amounts.

REFERENCES

- [1] L. BASS, *Functional zones in rat liver: the degree of overlap*, J. Theoret. Bio., 89 (1981), pp. 303–319.
- [2] L. BASS, A. J. BRACKEN AND R. VYBORNY, *Minimisation problems for implicit functionals defined by differential equations of liver kinetics*, J. Austral. Math. Soc., 25 (1984), pp. 538–562.
- [3] R.P. BRENT, *Algorithms for Minimization Without Derivatives*, Ser. on Automatic Computation, Prentice–Hall, New York, 1973.
- [4] E.A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, Tata McGraw–Hill, New Delhi, 1955, TMH edition 1972, pp. 57–61.
- [5] A.M. FINK, *Optimal control in liver kinetics*, J. Austral. Math. Soc., 27 (1986), pp. 361–369.
- [6] K. HOLMÅKER, *An optimal control problem in the study of liver kinetics*, J. Optim. Theory Appl., 48 (1986), pp. 289–302.
- [7] K. HOLMÅKER AND D. STEWART, *A class of optimization problems with noncompact constraints: General results and applications*, SIAM J. Control and Optim., 25 (1987), pp. 1032–1052.
- [8] K. STROMBERG, *Introduction to Classical Real Analysis*, Wadsworth, Inc., Belmont, CA, 1981, pp. 318–327.

ON THE LINEAR CONVERGENCE OF DESCENT METHODS FOR CONVEX ESSENTIALLY SMOOTH MINIMIZATION*

ZHI-QUAN LUO[†] AND PAUL TSENG[‡]

*Dedicated to those courageous people who, on June 4, 1989, sacrificed their lives in
Tiananmen Square, Beijing.*

Abstract. Consider the problem of minimizing, over a polyhedral set, the composition of an affine mapping with a strictly convex essentially smooth function. A general result on the linear convergence of descent methods for solving this problem is presented. By applying this result, the linear convergence of both the gradient projection algorithm of Goldstein and Levitin and Polyak, and a matrix splitting algorithm using regular splitting, is established. The results do not require that the cost function be strongly convex or that the optimal solution set be bounded. The key to the analysis lies in a new error bound for estimating the distance from a feasible point to the optimal solution set.

Key words. linear convergence, differentiable minimization, gradient projection, matrix splitting, coordinate descent

AMS(MOS) subject classifications. 49, 90

1. Introduction. Consider the problem of minimizing a strictly convex essentially smooth function subject to linear constraints. This problem contains a number of important optimization problems as special cases, including (strictly) convex quadratic programs, “ $x \ln(x)$ ” entropy minimization problems [Fri75], [Her80], [Jay82], [JoS84], [LaS81], [Pow88], and “ $-\ln(x)$ ” minimization problems [FiM68], [GMSTW86], [JoS84], [Son88]. A popular approach to solving this problem is to dualize the linear constraints to obtain a dual problem of minimizing, over a box, the composition of a strictly convex essentially smooth function with an affine mapping; then to apply a feasible descent method to solve the dual problem (see [Cen88], [CeL87], [CoP82], [Cry71], [Hil57], [Kru37], [LaS81], [LiP87], [MaD87], [MaD88a], [Tse90], [TsB87a], [TsB87b] and references therein). Popular choices for the descent method include a gradient projection algorithm of Goldstein [Gol64] and Levitin and Polyak [LeP65], the coordinate descent algorithm, and a matrix splitting algorithm using regular splitting [Kel65], [Man77], [OrR70], [Pan82].

An outstanding theoretical question concerns the *rate* of convergence of the iterates generated by the above solution approach. Most of the existing rate of convergence results require restrictive assumptions on the problem, such as that the cost function be strongly convex, which unfortunately do not hold in many practical situations. In fact, owing to the possible unboundedness of the optimal solution set, even the convergence of the iterates has been very difficult to establish (see [Che84], [LuT89a]). Nonetheless, by exploiting the special structure of the problem, it has been possible to

*Received by the editors June 4, 1990; accepted for publication (in revised form) February 25, 1991. This work was supported by U.S. Army Research Office contract DAAL03-86-K-0171 (Center for Intelligent Control Systems), by National Science Foundation grant NSF-DDM-8903385, and by a grant from the Science and Engineering Research Board of McMaster University.

[†]Room 225/CRL, Department of Electrical and Computer Engineering, McMaster University, Hamilton, Ontario, L8S 4L7, Canada.

[‡]Department of Mathematics, GN-50, University of Washington, Seattle, Washington 98195.

show *linear* convergence for two of the aforementioned algorithms: the gradient projection algorithm using small stepsizes (and under an additional Lipschitz assumption on the gradient) [BeG82] and the coordinate descent algorithm [LuT89b]. (Here and throughout, we mean by “linear convergence” the R -linear convergence in the sense of Ortega and Rheinboldt [OrR70].) In fact, the results for the gradient projection algorithm extend to variational inequality problems [BeG82].

In this paper, we extend the proof ideas and the results of [LuT89b] to a general class of feasible descent methods, including the aforementioned algorithms. In particular, we consider an extension of the above dual problem in which the constraint set is any polyhedral set, not just a box; we give general conditions for a feasible descent method to be linearly convergent when applied to solving this problem; and we show that the aforementioned algorithms (gradient projection, matrix splitting, etc.) satisfy these conditions and hence are linearly convergent when applied to solving this problem. The key to our analysis lies in a new bound for estimating the distance from a feasible point to the optimal solution set which, unlike many existing bounds, holds without requiring the cost function to be strongly convex.

We formally state our problem below. Let $f : \mathbb{R}^n \mapsto (-\infty, \infty]$ be a function of the form

$$(1.1) \quad f(x) = g(Ex) + \langle q, x \rangle, \quad \forall x,$$

where $g : \mathbb{R}^m \mapsto (-\infty, \infty]$ is some function, E is some $m \times n$ matrix (possibly with zero columns), and q is some vector in \mathbb{R}^n , the n -dimensional Euclidean space. In our notation, all vectors are column vectors and $\langle \cdot, \cdot \rangle$ denotes the usual Euclidean inner product. Notice that if q is in the row span of E , then f depends on x through Ex only. In general, however, this need not be the case.

We make the following standing assumptions regarding the function g .

ASSUMPTION 1.1. (a) The effective domain of g , denoted by C_g , is nonempty and open.

(b) g is strictly convex twice continuously differentiable on C_g .

(c) $g(t) \rightarrow \infty$ as t approaches any boundary point of C_g .

Assumption 1.1 implies that g is, in the terminology of Rockafellar [Roc70], a strictly convex *essentially smooth* function. Such a function has a number of interesting theoretical properties. For example, its conjugate function is also strictly convex essentially smooth (see [Roc70, Chap. 26]).

Let X be a polyhedral set in \mathbb{R}^n . Consider the following convex program associated with f and X :

$$(1.2) \quad \begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in X. \end{array}$$

We make the following standing assumptions regarding f and X .

ASSUMPTION 1.2. (a) The set of optimal solutions for (1.2), denoted by X^* , is nonempty.

(b) $\nabla^2 g(Ex^*)$ is positive definite for every $x^* \in X^*$.

Part (a) of Assumption 1.2 implies that the effective domain of f , denoted by C_f , makes a nonempty intersection with X and, moreover, f is bounded from below on X . It then follows from the special form of f (cf. (1.1)) and Assumption 1.1 that C_f is open, f is convex twice continuously differentiable on C_f , and $f(x) \rightarrow \infty$ as

x approaches any boundary point of C_f . Hence, f is convex essentially smooth but, unlike g , not necessarily strictly convex, since E may lack full column rank.

Part (b) of Assumption 1.2 states that g has a positive curvature on the image of X^* under the affine transformation $x \mapsto Ex$. This condition is guaranteed to hold if g has a positive curvature everywhere on C_g . (There are many important functions that satisfy this latter condition (in addition to Assumption 1.1), the most notable of which are the quadratic function, the exponential function, and the negative of the logarithm function.) Of course, if g is strongly convex and twice differentiable everywhere, then Assumptions 1.1 and 1.2 (b) hold automatically. We remark that the twice differentiability of g is not necessary for our results to hold, but it makes for a simpler statement of the assumptions. In general it suffices that g be differentiable on C_g and that ∇g be “locally” strongly monotone and Lipschitz continuous around Ex^* for all $x^* \in X^*$.

The problem (1.2) contains a number of important problems as special cases. For example, if E is the null matrix, then (1.2) reduces to a linear program. If g is a strictly convex quadratic function, then (1.2) reduces to the symmetric monotone linear complementarity problem [Man77], [LiP87] (also see §5). If g is the function given by $g(t) = -\sum_j \ln(t_j)$ for all $t \in (0, \infty)^m$ and $g(t) = \infty$ otherwise, where $\ln(\cdot)$ denotes the natural logarithm, and X is the nonnegative orthant in \mathbb{R}^n , then (1.2) reduces to the Lagrangian dual of a certain linearly constrained logarithmic penalty problem (see, e.g., [CeL87]).

For any $x \in \mathbb{R}^n$, let $[x]^+$ denote the orthogonal projection of x onto X , i.e.,

$$[x]^+ = \arg \min_{y \in X} \|x - y\|,$$

where $\|\cdot\|$ denotes the usual Euclidean norm (i.e., $\|x\| = \sqrt{\langle x, x \rangle}$ for all x). (Our notation for the projection operator is somewhat unconventional, but it has the advantage of simplicity.) Since C_f is nonempty, and f is differentiable on C_f , it is easily seen from the Kuhn–Tucker conditions for (1.2) that X^* comprises all $x \in X \cap C_f$ for which the orthogonal projection of $x - \nabla f(x)$ onto X is x itself, i.e.,

$$(1.3) \quad X^* = \{ x \in \mathbb{R}^n \mid x = [x - \nabla f(x)]^+ \}.$$

Note that since both f and X are closed and convex, then so is X^* (in fact, X^* is a polyhedral set as we show in Corollary 2.1). However, X^* may be unbounded.

This paper proceeds as follows. In §2 we prove a new bound on the distance to X^* from a feasible point near X^* . In §3 we use this bound to establish general conditions under which a sequence of feasible points converge at least linearly to an element of X^* . In §§4 and 5 we show that the sequence of iterates generated by either the gradient projection algorithm or the matrix splitting algorithm using regular splitting satisfy the convergence conditions outlined in §3. In §6 we give our conclusion and discuss possible extensions.

We adopt the following notation throughout. For any $k \times l$ matrix A , we denote by A^T the transpose of A , by $\|A\|$ the matrix norm of A induced by the vector norm $\|\cdot\|$ (i.e., $\|A\| = \max_{\|x\|=1} \|Ax\|$), by A_i the i th row of A and, for any nonempty $I \subseteq \{1, \dots, k\}$, by A_I the submatrix of A obtained by removing all rows $i \notin I$ of A . Analogously, for any vector $x \in \mathbb{R}^k$, we denote by x_i the i th coordinate of x , and, for any nonempty subset $I \subseteq \{1, \dots, k\}$, by x_I the vector with components x_i , $i \in I$ (with the x_i ’s arranged in the same order as in x).

2. A new error bound. In this section we prove a key result that, for all $x \in X \cap C_f$ sufficiently close to the optimal solution set X^* , the distance from x to X^* is of the order $\|x - [x - \nabla f(x)]\|$. This result will be used in the rate of convergence analysis of §3.

We first need the following lemma which says that the affine mapping $x \mapsto Ex$ is invariant over X^* . This lemma is a simple consequence of the strict convexity of g .

LEMMA 2.1. *There exists a $t^* \in \mathbb{R}^m$ such that*

$$Ex^* = t^*, \quad \forall x^* \in X^*.$$

Proof. For any $x^* \in X^*$ and $y^* \in X^*$, we have by the convexity of X^* that $(x^* + y^*)/2 \in X^*$. Then, $f(x^*) = f(y^*) = f((x^* + y^*)/2)$, so (1.1) yields $g((Ex^* + Ey^*)/2) = (g(Ex^*) + g(Ey^*))/2$. Since both $g(Ex^*)$ and $g(Ey^*)$ are finite, so that $Ex^* \in C_g$ and $Ey^* \in C_g$, this together with the strict convexity of g on C_g implies $Ex^* = Ey^*$. \square

As an immediate corollary of Lemma 2.1, we have the following interesting characterization of X^* .

COROLLARY 2.1. *X^* is a polyhedral set.*

Proof. Fix any $x^* \in X^*$ and consider the polyhedral set

$$\bar{X} = \{ x \in X \mid Ex = t^*, \langle q, x - x^* \rangle = 0 \}.$$

By using (1.1) and Lemma 2.1, we see that x is an element of \bar{X} if and only if $x \in X$ and $f(x) = f(x^*)$. Thus, $\bar{X} = X^*$. \square

By using the observation (cf. (1.1))

$$(2.1) \quad \nabla f(x) = E^T \nabla g(Ex) + q, \quad \forall x \in C_f,$$

we have that ∇f is invariant over X^* . In fact, it is easily seen from Lemma 2.1 that

$$(2.2) \quad \nabla f(x^*) = d^*, \quad \forall x^* \in X^*,$$

where we denote

$$(2.3) \quad d^* = E^T \nabla g(t^*) + q.$$

The above invariant property of ∇f on X^* is quite well known (see, e.g., [Man88]) and in fact holds for more general convex programs.

Since $\nabla^2 g(t^*)$ is positive definite (cf. Assumption 1.2 (b) and Lemma 2.1), it follows from the continuity property of $\nabla^2 g$ (Assumption 1.1 (b)) that $\nabla^2 g$ is positive definite in some open neighborhood of t^* . This in turn implies that g is strongly convex near t^* , i.e., there exist a positive scalar $\sigma > 0$ and a closed ball $\mathcal{U}^* \subseteq C_g$ containing t^* such that

$$(2.4) \quad g(z) - g(y) - \langle \nabla g(y), z - y \rangle \geq \sigma \|z - y\|^2, \quad \forall z \in \mathcal{U}^*, \forall y \in \mathcal{U}^*.$$

By interchanging the role of y with that of z in (2.4) and adding the resulting relation to (2.4), we also obtain

$$(2.5) \quad \langle \nabla g(z) - \nabla g(y), z - y \rangle \geq 2\sigma \|z - y\|^2, \quad \forall z \in \mathcal{U}^*, \forall y \in \mathcal{U}^*.$$

Since $\nabla^2 g$ is continuous on \mathcal{U}^* (cf. Assumption 1.1 (b)), so it is bounded there, then ∇g is Lipschitz continuous on \mathcal{U}^* , i.e., there exists a scalar $\rho > 0$ such that

$$(2.6) \quad \|\nabla g(z) - \nabla g(y)\| \leq \rho \|z - y\|, \quad \forall z \in \mathcal{U}^*, \forall y \in \mathcal{U}^*.$$

We next state a lemma on the Lipschitz continuity of the solution set of a linear system as a multifunction of the right-hand side. This lemma, originally due to [Hof52] (also see [Rob73], [MaS87]), will be used in the proof of Lemmas 2.4, 2.6, and 3.1.

LEMMA 2.2. *Let B be a $k \times l$ matrix, let C be an $h \times l$ matrix, and let d be a vector in \mathbb{R}^h . There exists a scalar $\theta > 0$ depending on B and C only such that, for any \bar{x} satisfying $C\bar{x} \geq d$ and any $e \in \mathbb{R}^k$ such that the linear system $By = e$, $Cy \geq d$ is consistent, there is a point \bar{y} satisfying $B\bar{y} = e$, $C\bar{y} \geq d$ and $\|\bar{x} - \bar{y}\| \leq \theta \|B\bar{x} - e\|$.*

By using Lemma 2.2 and Assumption 1.1, we can show the following technical fact.

LEMMA 2.3. *For any $\zeta \in \mathbb{R}$, the set $\{Ex \mid x \in X, f(x) \leq \zeta\}$ is a compact subset of C_g .*

Proof. For the proof see [Tse89, Lem. 9.1].

Since X is a polyhedral set, we can for convenience express it as $X = \{x \in \mathbb{R}^n \mid Ax \geq b\}$, for some $k \times n$ matrix A and some $b \in \mathbb{R}^k$. Then, for any $x \in X \cap C_f$, the vector $[x - \nabla f(x)]^+$ is the unique vector z which, together with some multiplier vector $\lambda \in \mathbb{R}^k$, satisfies the Kuhn–Tucker conditions

$$(2.7) \quad z - x + \nabla f(x) - A^T \lambda = 0, \quad Az \geq b, \quad \lambda \geq 0,$$

$$(2.8) \quad \lambda_i = 0, \quad \forall i \notin I(x), \quad A_i z = b_i, \quad \forall i \in I(x),$$

where we denote

$$I(x) = \{i \in \{1, \dots, k\} \mid A_i[x - \nabla f(x)]^+ = b_i\}.$$

We say that an $I \subseteq \{1, \dots, k\}$ is *active* at a vector $x \in X \cap C_f$ if $z = [x - \nabla f(x)]^+$ together with some $\lambda \in \mathbb{R}^k$ satisfies (2.7) and

$$(2.9) \quad \lambda_i = 0, \quad \forall i \notin I, \quad A_i z = b_i, \quad \forall i \in I.$$

(By (2.8), $I(x)$ is active at x for all $x \in X \cap C_f$).

We next have the following lemma, which roughly says that if $x \in X$ is sufficiently close to X^* , then those constraint indices that are active at x are also active at some element of X^* .

LEMMA 2.4. *For any $\zeta \in \mathbb{R}$, there exists an $\epsilon > 0$ such that, for any $x \in X$ with $f(x) \leq \zeta$ and $\|x - [x - \nabla f(x)]^+\| \leq \epsilon$, $I(x)$ is active at some $x^* \in X^*$.*

Proof. We argue by contradiction. If the claim does not hold, then for some $\zeta \in \mathbb{R}$, there would exist an $I \subseteq \{1, \dots, k\}$ and a sequence of vectors $\{x^r\}$ in X satisfying $f(x^r) \leq \zeta$ for all r , $x^r - z^r \rightarrow 0$, where we let $z^r = [x^r - \nabla f(x^r)]^+$ for all r , and $I(x^r) = I$ for all r , and yet there is no $x^* \in X^*$ for which I is active at x^* .

Since $\{f(x^r)\}$ is bounded, it follows from Lemma 2.3 that $\{Ex^r\}$ lies in a compact subset of C_g . Let t^∞ be any such cluster point of $\{Ex^r\}$ (so $t^\infty \in C_g$) and let \mathcal{R} be a subsequence of $\{0, 1, \dots\}$ such that

$$(2.10) \quad \{Ex^r\}_{\mathcal{R}} \rightarrow t^\infty.$$

We show below that t^∞ is equal to t^* .

Since $t^\infty \in C_g$ so ∇g is continuous in an open set around t^∞ , we obtain from (2.10) (and using the fact $\nabla f(x^r) = E^T \nabla g(Ex^r) + q$ for all r) that

$$(2.11) \quad \{\nabla f(x^r)\}_{\mathcal{R}} \rightarrow E^T \nabla g(t^\infty) + q.$$

For each $r \in \mathcal{R}$, consider the following linear system in x , z , and λ :

$$\begin{aligned} x - z + A^T \lambda &= \nabla f(x^r), & Ex &= Ex^r, & x - z &= x^r - z^r, \\ Az &\geq b, & \lambda &\geq 0, & A_i z &= b_i, \quad \forall i \in I, & \lambda_i &= 0, \quad \forall i \notin I. \end{aligned}$$

The above system is consistent since, by $I(x^r) = I$ and (2.7)–(2.8), (x^r, z^r) together with some $\lambda^r \in \mathbb{R}^k$ is a solution of it. Then, by Lemma 2.2, it has a solution $(\hat{x}^r, \hat{z}^r, \hat{\lambda}^r)$ whose size is bounded by some constant (depending on A and E only) times the size of the right-hand side. Since the right-hand side of the above system, by $z^r - x^r \rightarrow 0$ and (2.10)–(2.11), is bounded as $r \rightarrow \infty$, $r \in \mathcal{R}$, then $\{(\hat{x}^r, \hat{z}^r, \hat{\lambda}^r)\}_{\mathcal{R}}$ is bounded. Moreover, every one of its cluster points, say $(x^\infty, z^\infty, \lambda^\infty)$, satisfies

$$\begin{aligned} x^\infty - z^\infty + A^T \lambda^\infty &= E^T \nabla g(t^\infty) + q, & Ex^\infty &= t^\infty, & x^\infty - z^\infty &= 0, \\ Az^\infty &\geq b, & \lambda^\infty &\geq 0, & A_i z^\infty &= b_i, \quad \forall i \in I, & \lambda_i^\infty &= 0, \quad \forall i \notin I. \end{aligned}$$

This shows $x^\infty = [x^\infty - \nabla f(x^\infty)]^+$ (cf. (2.7), (2.8), and $\nabla f(x^\infty) = E^T \nabla g(Ex^\infty) + q$), so $x^\infty \in X^*$ (cf. (1.3)) and, by Lemma 2.1, $t^\infty = t^*$. Moreover, I is active at x^∞ (cf. (2.7) and (2.9)), so a contradiction is established. \square

Also, the proof of Lemma 2.4 shows the following lemma.

LEMMA 2.5. *For any $\zeta \in \mathbb{R}$ and any $\eta > 0$, there exists an $\epsilon' > 0$ such that $\|Ex - t^*\| \leq \eta$ for all $x \in X$ with $f(x) \leq \zeta$ and $\|x - [x - \nabla f(x)]^+\| \leq \epsilon'$.*

By using Lemmas 2.2, 2.4, and 2.5, we can prove the following intermediate lemma.

LEMMA 2.6. *For any $\zeta \in \mathbb{R}$, there exist scalars $\delta > 0$ and $\omega > 0$ such that, for any $x \in X$ with $f(x) \leq \zeta$ and $\|x - [x - \nabla f(x)]^+\| \leq \delta$, the following hold:*

- (a) $Ex \in \mathcal{U}^*$.
- (b) *There exists a $\lambda \in [0, \infty)^k$ satisfying*

$$(A_I)^T \lambda_I = z - x + \nabla f(x),$$

and an $x^ \in X^*$ and a $\lambda^* \in [0, \infty)^k$ satisfying*

$$(A_I)^T \lambda_I^* = \nabla f(x^*), \quad A_I x^* = b_I,$$

$$\|(x, \lambda) - (x^*, \lambda^*)\| \leq \omega (\|Ex - t^*\| + \|x - z\|),$$

where we let $I = I(x)$ and $z = [x - \nabla f(x)]^+$.

Proof. Fix any $\zeta \in \mathbb{R}$. By Lemma 2.5, there exists some $\epsilon' > 0$ such that $Ex \in \mathcal{U}^*$ for all $x \in X$ satisfying $f(x) \leq \zeta$ and $\|x - [x - \nabla f(x)]^+\| \leq \epsilon'$. Choose δ to be the minimum of this ϵ' and the ϵ given in Lemma 2.4.

Consider any $x \in X$ satisfying the hypothesis of the lemma (with the above choice of δ), and let $z = [x - \nabla f(x)]^+$. Then, by (2.7) and (2.8), there exists some $\lambda \in \mathbb{R}^k$ satisfying, together with x and z ,

$$A^T \lambda = z - x + \nabla f(x),$$

$$Az \geq b, \quad \lambda \geq 0, \quad \lambda_i = 0, \quad \forall i \notin I(x), \quad A_i z = b_i, \quad \forall i \in I(x).$$

By Lemma 2.4, there exists an $x^* \in X^*$ such that $I(x)$ is active at x^* , so, by Lemma 2.1 and (2.2), the following linear system in (x^*, z^*, λ^*) :

$$A^T \lambda^* = d^*, \quad Ex^* = t^*, \quad x^* - z^* = 0,$$

$$Az^* \geq b, \quad \lambda^* \geq 0, \quad \lambda_i^* = 0, \quad \forall i \notin I(x), \quad A_i z^* = b_i, \quad \forall i \in I(x),$$

is consistent. Moreover, every solution (x^*, z^*, λ^*) of this linear system satisfies $x^* \in X^*$. Upon comparing the above two systems, we see that, by Lemma 2.2, there exists a solution (x^*, z^*, λ^*) to the second system such that

$$\|(x^*, z^*, \lambda^*) - (x, z, \lambda)\| \leq \theta(\|z - x + \nabla f(x) - d^*\| + \|Ex - t^*\| + \|x - z\|),$$

where θ is some scalar depending on A and E only. Since our choice of δ also implies that $Ex \in \mathcal{U}^*$, (2.1)–(2.3) and the Lipschitz condition (2.6) yield $\|\nabla f(x) - d^*\| = \|E^T \nabla g(Ex) - E^T \nabla g(t^*)\| \leq \rho \|E^T\| \|Ex - t^*\|$. Hence the above relation implies

$$\|(x^*, z^*, \lambda^*) - (x, z, \lambda)\| \leq \theta(2\|x - z\| + (\rho\|E^T\| + 1)\|Ex - t^*\|). \quad \square$$

For any $x \in \mathbb{R}^n$, let $\phi(x)$ denote the Euclidean distance from x to X^* , i.e.,

$$\phi(x) = \min_{x^* \in X^*} \|x - x^*\|.$$

By using Lemmas 2.1 and 2.6, we can establish the main result of this section, which roughly says that, for all $x \in X \cap C_f$ sufficiently close to X^* , $\phi(x)$ can be bounded from above by the norm of the residual $x - [x - \nabla f(x)]^+$.

THEOREM 2.1. *For any $\zeta \in \mathbb{R}$, there exist scalars $\tau > 0$ and $\delta > 0$ such that*

$$(2.12) \quad \phi(x) \leq \tau \|x - [x - \nabla f(x)]^+\|,$$

for all $x \in X$ with $f(x) \leq \zeta$ and $\|x - [x - \nabla f(x)]^+\| \leq \delta$.

Proof. Fix any $\zeta \in \mathbb{R}$ and let δ and ω be the corresponding scalars given in Lemma 2.6.

Consider any $x \in X$ satisfying the hypothesis of the theorem (with the above choice of δ), and let $z = [x - \nabla f(x)]^+$, $I = I(x)$. Then,

$$(2.13) \quad A_I z = b_I,$$

and, by Lemma 2.6, $Ex \in \mathcal{U}^*$ and there exists a $\lambda \in [0, \infty)^k$ satisfying

$$(2.14) \quad (A_I)^T \lambda_I = z - x + \nabla f(x),$$

and an $x^* \in X^*$ and a $\lambda^* \in [0, \infty)^k$ satisfying

$$(2.15) \quad (A_I)^T \lambda_I^* = \nabla f(x^*), \quad A_I x^* = b_I,$$

$$(2.16) \quad \|(x, \lambda) - (x^*, \lambda^*)\| \leq \omega(\|Ex - t^*\| + \|x - z\|).$$

Also, since $Ex \in \mathcal{U}^*$ and $t^* \in \mathcal{U}^*$, we have from (2.5) that

$$(2.17) \quad 2\sigma \|Ex - t^*\|^2 \leq \langle Ex - t^*, \nabla g(Ex) - \nabla g(t^*) \rangle.$$

We claim that (2.13)–(2.17) are sufficient to establish (2.12). To see this, note that $Ex^* = t^*$ (cf. Lemma 2.1) and $\nabla f(x) - \nabla f(x^*) = E^T \nabla g(Ex) - E^T \nabla g(Ex^*)$ (cf. (2.1)), so (2.17) and (2.13)–(2.15) yield

$$\begin{aligned} 2\sigma \|Ex - t^*\|^2 &\leq \langle Ex - t^*, \nabla g(Ex) - \nabla g(t^*) \rangle \\ &= \langle x - x^*, \nabla f(x) - \nabla f(x^*) \rangle \\ &= \langle x - x^*, (A_I)^T \lambda_I - z + x - (A_I)^T \lambda_I^* \rangle \\ &= \langle A_I(x - x^*), \lambda_I - \lambda_I^* \rangle + \langle x - x^*, x - z \rangle \\ &= \langle A_I(x - z), \lambda_I - \lambda_I^* \rangle + \langle x - x^*, x - z \rangle \\ &= O(\|x - z\|(\|\lambda - \lambda^*\| + \|x - x^*\|)), \end{aligned}$$

where for convenience we use the notation $\alpha = O(\beta)$ to indicate that $\alpha \leq \kappa\beta$ for some scalar $\kappa > 0$ depending on ζ and the problem data only. Combining the above relation with (2.16) then gives

$$\begin{aligned} \|x - x^*\|^2 &\leq \omega^2 (\|Ex - t^*\| + \|x - z\|)^2 \\ &\leq 2\omega^2 (\|Ex - t^*\|^2 + \|x - z\|^2) \\ &= O(\|x - z\|(\|\lambda - \lambda^*\| + \|x - x^*\|) + \|x - z\|^2) \\ &= O(\|x - z\|(\|Ex - t^*\| + \|x - z\|)), \end{aligned}$$

where the last step follows from using (2.16). Hence $\|Ex - t^*\|^2$, which is clearly $O(\|x - x^*\|^2)$, must be $O(\|x - z\|(\|Ex - t^*\| + \|x - z\|))$, i.e., there exists a scalar $\kappa > 0$ (depending on ζ and the problem data only) such that

$$\|Ex - t^*\|^2 \leq \kappa (\|x - z\| \|Ex - t^*\| + \|x - z\|^2).$$

This is a quadratic inequality of the form $a^2 \leq \kappa(ab + b^2)$, which implies $a \leq (\kappa + \sqrt{\kappa^2 + 4\kappa})b/2$. Hence we obtain that

$$\|Ex - t^*\| \leq (\kappa + \sqrt{\kappa^2 + 4\kappa})\|x - z\|/2,$$

which when combined with (2.16) shows $\|x - x^*\| = O(\|x - z\|)$. Since $x^* \in X^*$ so clearly $\phi(x) \leq \|x - x^*\|$, this then completes our proof. \square

We remark that computable error bounds like the one given in Theorem 2.1 have been quite well studied. In fact, a bound identical to that given in Theorem 2.1 was proposed by Pang for the special case where f is strongly convex [Pan87]. Alternative bounds have also been proposed for strongly convex programs [MaD88b] and for monotone linear complementarity problems [MaS86]. However, it is unclear whether these alternative bounds are useful for analyzing the rate of convergence of algorithms. On the other hand, note that the bound in Theorem 2.1 holds only locally, and it would be interesting to see whether this bound can be extended to hold globally on $X \cap C_f$.

3. A general linear convergence result. In this section we give general conditions for a sequence of points in $X \cap C_f$ to converge at least linearly to an optimal

solution of (1.2) or, equivalently, to an element of X^* . This result, based in large part on the error bound developed in §2, will be used in §§4 and 5 to establish the linear convergence of a gradient projection algorithm and a matrix splitting algorithm.

We first state the following simple lemma.

LEMMA 3.1. *Let $\{x^r\}$ be a sequence of vectors in $X \cap C_f$ satisfying*

$$(3.1) \quad \|x^r - x^{r+1}\|^2 \leq \kappa_1(f(x^r) - f(x^{r+1})), \quad \forall r \geq r_0,$$

for some scalar $\kappa_1 > 0$. If $f(x^r)$ converges linearly, then $\{x^r\}$ also converges linearly.

(Recall that here “linear convergence” means R -linear convergence in the sense of Ortega and Rheinboldt [OrR70].)

The general linear convergence result is the following theorem.

THEOREM 3.1. *Let v^* denote the optimal value of (1.2). Let $\{x^r\}$ be a sequence of vectors in $X \cap C_f$ satisfying the following two conditions:*

$$(3.2) \quad f(x^r) - v^* \leq \kappa_2 \phi(x^r)^2, \quad \forall r \geq r_0,$$

$$(3.3) \quad \|x^r - [x^r - \nabla f(x^r)]^+\|^2 \leq \kappa_3(f(x^r) - f(x^{r+1})), \quad \forall r \geq r_0,$$

where κ_2, κ_3 and r_0 are some positive scalars. Then, $\{f(x^r)\}$ converges at least linearly to v^* . If, in addition, (3.1) holds, then $\{x^r\}$ converges at least linearly to an element of X^* .

Proof. By (3.3), $\{f(x^r)\}$ is monotonically decreasing. Since f is also bounded from below on X (cf. Assumption 1.2 (a)), then $f(x^r) - f(x^{r+1}) \rightarrow 0$, so (3.3) yields $\|x^r - [x^r - \nabla f(x^r)]^+\| \rightarrow 0$. Hence, by Theorem 2.1, there exist scalars $\tau > 0$ and $r_1 \geq r_0$ such that

$$(3.4) \quad \phi(x^r) \leq \tau \|x^r - [x^r - \nabla f(x^r)]^+\|, \quad \forall r \geq r_1.$$

By combining (3.2), (3.3), and (3.4), we obtain that, for each $r \geq r_1$, there holds

$$(3.5) \quad \begin{aligned} f(x^r) - v^* &\leq \kappa_2 \phi(x^r)^2 \\ &\leq \kappa_2 \tau^2 \|x^r - [x^r - \nabla f(x^r)]^+\|^2 \\ &\leq \kappa_2 \kappa_3 \tau^2 (f(x^r) - f(x^{r+1})). \end{aligned}$$

Upon rearranging terms in (3.5), we then obtain

$$f(x^{r+1}) - v^* \leq \left(1 - \frac{1}{\kappa_2 \kappa_3 \tau^2}\right) (f(x^r) - v^*),$$

so $\{f(x^r)\}$ converges at least linearly to v^* . If (3.1) holds, then it follows from Lemma 3.1 that $\{x^r\}$ converges at least linearly. Let x^∞ denote the limit point of $\{x^r\}$. Then, $x^\infty \in X$ (since X is closed) and, by the lower semicontinuity of f , $f(x^\infty) \leq v^*$. Therefore $x^\infty \in X^*$. \square

(Roughly speaking, (3.2) says that the difference in cost between an iterate and its nearest optimal solution should grow *at most* quadratically in the distance between them; and (3.3) says that the decrease in the cost at each iteration should grow *at least* quadratically in the “residual” at the current iterate.)

It turns out that, for our applications (see §§4 and 5), (3.1) and (3.3) are relatively easy to verify. The difficulty lies in verifying that (3.2) holds. To help us with this

endeavor, we develop below, by using Lemmas 2.2, 2.3, and 2.5, a sufficient condition for (3.2) to hold. This condition, although more restrictive than (3.2), is much easier to verify for the algorithms considered in this paper.

LEMMA 3.2.. *Suppose that $\{x^r\}$ satisfies (3.1) and (3.3) for some scalars κ_1, κ_3 , and r_0 , and*

$$(3.6) \quad x^{r+1} = [x^r - \alpha^r \nabla f(x^r) + e^r]^+, \quad \forall r \geq r_1,$$

for some index r_1 , some bounded sequence of scalars $\{\alpha^r\}$ bounded away from zero and some sequence of n -vectors $e^r \rightarrow 0$. Then $\{x^r\}$ satisfies (3.2) for some scalar κ_2 (possibly with a different value for r_0).

Proof. Since f is bounded from below on X (cf. Assumption 1.2 (a)) and (3.1) and (3.3) hold, then

$$(3.7) \quad x^r - x^{r-1} \rightarrow 0,$$

$$(3.8) \quad x^r - [x^r - \nabla f(x^r)]^+ \rightarrow 0.$$

We claim that there exists an index $r_2 \geq r_1$ such that

$$(3.9) \quad \langle \nabla f(x^*), x^r - x^* \rangle = 0, \quad \forall x^* \in X^*,$$

for all $r \geq r_2$. To see this, let X be expressed as $X = \{x \in \mathbb{R}^n \mid Ax \geq b\}$, for some $k \times n$ matrix A and some k -vector b , and, for every $r \geq r_1$, let I^r denote the set of indices $i \in \{1, \dots, k\}$ such that $A_i x^r = b_i$. For any $I \subseteq \{1, \dots, k\}$, define the index set $\mathcal{R}_I = \{r \in \{1, 2, \dots\} \mid I^r = I\}$. It suffices to show that, for any I with \mathcal{R}_I infinite, (3.9) holds for all $r \in \mathcal{R}_I$ sufficiently large. We show this below.

Fix any I for which \mathcal{R}_I is infinite. Our argument will follow closely the proof of Lemma 2.4. Since $\{f(x^r)\}$ is monotonically decreasing (cf. (3.1)), then it is bounded. Hence, by Lemma 2.3, $\{Ex^r\}$ lies in a compact subset of C_g . Let t^∞ be any cluster point of $\{Ex^{r-1}\}_{\mathcal{R}_I}$ (so $t^\infty \in C_g$) and let \mathcal{R}' be any subsequence of \mathcal{R}_I such that

$$(3.10) \quad \{Ex^{r-1}\}_{\mathcal{R}'} \rightarrow t^\infty.$$

Let $d^\infty = E^T \nabla g(t^\infty) + q$. Then, since $t^\infty \in C_g$ so ∇g is continuous in an open set around t^∞ , we obtain from (3.10) (and using the fact $\nabla f(x^r) = E^T \nabla g(Ex^r) + q$ for all r) that

$$(3.11) \quad \{\nabla f(x^{r-1})\}_{\mathcal{R}'} \rightarrow d^\infty.$$

For each $r \in \mathcal{R}'$, consider the following linear system in x, z , and λ :

$$\begin{aligned} x - z + A^T \lambda &= \alpha^{r-1} \nabla f(x^{r-1}) - e^{r-1}, & Ex &= Ex^{r-1}, & x - z &= x^{r-1} - x^r, \\ Az &\geq b, & \lambda &\geq 0, & A_i z &= b_i, \quad \forall i \in I, & \lambda_i &= 0, \quad \forall i \notin I. \end{aligned}$$

The above system is consistent since, by $I^r = I$ and (3.6), it is satisfied by $x = x^{r-1}$, $z = x^r$, and some λ in \mathbb{R}^k (cf. (2.7)–(2.8)). Then, by Lemma 2.2, it has a solution $(\hat{x}^r, \hat{z}^r, \hat{\lambda}^r)$ whose size is bounded by some constant (depending on A and E only) times the size of the right-hand side. Since the right-hand side of the above system

is bounded as $r \rightarrow \infty$, $r \in \mathcal{R}'$ (cf. (3.7), (3.10)–(3.11), $e^r \rightarrow 0$ and the boundedness of $\{\alpha^r\}$), we have that $\{(\hat{x}^r, \hat{z}^r, \hat{\lambda}^r)\}_{\mathcal{R}'}$ is also bounded. Moreover, every one of its cluster points, say $(x^\infty, z^\infty, \lambda^\infty)$, satisfies together with some $\alpha^\infty > 0$ the following conditions (cf. (3.7), (3.10)–(3.11), $e^r \rightarrow 0$, and the boundedness hypothesis on $\{\alpha^r\}$)

$$(3.12) \quad x^\infty - z^\infty + A^T \lambda^\infty = \alpha^\infty d^\infty, \quad E x^\infty = t^\infty, \quad x^\infty - z^\infty = 0,$$

$$(3.13) \quad A z^\infty \geq b, \quad \lambda^\infty \geq 0, \quad A_i z^\infty = b_i, \quad \forall i \in I, \quad \lambda_i^\infty = 0, \quad \forall i \notin I.$$

This shows $x^\infty = [x^\infty - \alpha^\infty \nabla f(x^\infty)]^+$ (cf. (2.7)–(2.8), $\nabla f(x^\infty) = E^T \nabla g(E x^\infty) + q$ and the definition of d^∞) so, by (1.3), $x^\infty \in X^*$. Fix any $r \in \mathcal{R}_I$. From (3.12)–(3.13) we also have that $A_I x^\infty = b_I$ and $\alpha^\infty \nabla f(x^\infty) = (A_I)^T \lambda_I^\infty$. Since $A_I x^r = b_I$ (cf. $I = I^r$), we thus obtain

$$(3.14) \quad \langle \nabla f(x^\infty), x^r - x^\infty \rangle = \frac{1}{\alpha^\infty} \langle \lambda_I^\infty, A_I(x^r - x^\infty) \rangle = 0.$$

Since x^∞ belongs to the convex set X^* and f is constant on X^* , then we must also have

$$\langle \nabla f(x^\infty), x^* - x^\infty \rangle = 0, \quad \forall x^* \in X^*,$$

which together with (3.14) and the invariance of ∇f on X^* (cf. (2.2)) proves (3.9).

Since $f(x^r) \leq f(x^0)$ for all r (cf. (3.3)) and (3.8) holds, then Lemma 2.5 implies $E x^r \rightarrow t^*$, so there exists an index $r_3 \geq r_2$ such that $E x^r \in \mathcal{U}^*$ for all $r \geq r_3$. Fix any $r \geq r_3$ and let x^* be an element of X^* with $\|x^r - x^*\| = \phi(x^r)$. By the Mean Value Theorem, there exists a ξ lying on the line segment joining x^r with $x^* \in X^*$ such that $f(x^r) - f(x^*) = \langle \nabla f(\xi), x^r - x^* \rangle$. This combined with (3.9) then yields

$$\begin{aligned} f(x^r) - f(x^*) &= \langle \nabla f(\xi), x^r - x^* \rangle \\ &= \langle \nabla f(\xi) - \nabla f(x^*), x^r - x^* \rangle + \langle \nabla f(x^*), x^r - x^* \rangle \\ &= \langle \nabla g(E\xi) - \nabla g(E x^*), E x^r - E x^* \rangle + \langle \nabla f(x^*), x^r - x^* \rangle \\ &\leq \rho \|E\|^2 \|x^r - x^*\|^2 + \langle \nabla f(x^*), x^r - x^* \rangle \\ &= \rho \|E\|^2 \|x^r - x^*\|^2, \end{aligned}$$

where the third equality follows from (2.1) and the inequality follows from the Lipschitz condition (2.6) (recall that $E x^r \in \mathcal{U}^*$ and $E x^* = t^* \in \mathcal{U}^*$). Since $\|x^r - x^*\| = \phi(x^r)$, this shows that (3.2) holds (with κ_2 and r_0 set to, respectively, $\rho \|E\|^2$ and r_3). \square

A few remarks about (3.6) are in order. Condition (3.6) roughly says that the iterates $\{x^r\}$ should eventually identify those constraints representing X that are active at some optimal solution. To see why this helps us to show (3.2), consider the case where X is simply a box (i.e., bound constraints). In this case, (3.6) translates to say that, for all r sufficiently large, those coordinates x_i^r of x^r for which $d_i^* > 0$ (respectively, $d_i^* < 0$) become fixed at the upper (respectively, lower) bound of x_i , which is also the bound that is active for any optimal solution. Then, it follows that, for each such r , there holds $\langle d^*, x^r - x^* \rangle = 0$, for all $x^* \in X^*$ (compare with (3.9)) from which (3.2) readily follows. The scalars $\{\alpha^r\}$ can be thought of as stepsizes and are introduced to model algorithms which incorporate stepsizes into their iterations (such as the gradient projection algorithm and algorithms that employ line search steps). The vectors $\{e^r\}$ carry no meaning in themselves and are introduced mainly

as a convenient tool to simplify the analysis. However, that e^r appears *inside* the projection operator $[\cdot]^+$ is crucial, for otherwise the above constraint identification property would not hold. As we shall see, (3.6) can be used to model a fairly diverse class of algorithms by suitably choosing $\{\alpha^r\}$ and $\{e^r\}$.

By using Lemma 3.2, we immediately obtain the following useful corollary of Theorem 3.1.

COROLLARY 3.1. *Suppose that $\{x^r\}$ satisfies (3.1), (3.3), and (3.6) for some scalars $\kappa_1, \kappa_3, r_0, r_1$, some bounded sequence of scalars $\{\alpha^r\}$ bounded away from zero and some sequence of n -vectors $e^r \rightarrow 0$. Then $\{f(x^r)\}$ converges at least linearly to the optimal value of (1.2) and $\{x^r\}$ converges at least linearly to an element of X^* .*

4. Linear convergence of a gradient projection algorithm. In this section, we make (in addition to Assumptions 1.1 and 1.2) the following assumptions on f .

ASSUMPTION 4.1. (a) $C_f = \mathbb{R}^n$.

(b) ∇f is Lipschitz continuous on \mathbb{R}^n , that is,

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|, \quad \forall x, \forall y,$$

where L is the Lipschitz constant.

Consider the following algorithm of Goldstein [Gol64] and of Levitin and Polyak [LeP65] applied to solve this special case of (1.2).

GRADIENT PROJECTION ALGORITHM.

Let $\sigma \in (0, 1)$, $\underline{\alpha} \in (0, 2(1 - \sigma)/L)$ and $\bar{\alpha} \geq \underline{\alpha}$ be given scalars. At the r th iteration, we are given an $x^r \in X$ (with x^0 chosen arbitrarily) and we compute a new iterate x^{r+1} in X according to

$$(4.1) \quad x^{r+1} = [x^r - \alpha^r \nabla f(x^r)]^+,$$

where α^r is any scalar in the interval $[\underline{\alpha}, \bar{\alpha}]$ such that x^{r+1} given by (4.1) satisfies

$$(4.2) \quad f(x^r) - f(x^{r+1}) \geq \frac{\sigma}{\alpha^r} \|x^r - x^{r+1}\|^2.$$

We remark that the restriction (4.2) is fairly mild and is satisfied by a number of well-known stepsize rules for the gradient projection algorithm, including (i) the rule given by [Gol64] and [LeP65]:

$$\epsilon \leq \alpha^r \leq 2/L - \epsilon,$$

with $\epsilon \in (0, 1/L)$ a given scalar; (ii) the Armijo-like rule given by [Ber76] (also see [Dun81]) in which α^r is the *largest* nonnegative power of a given scalar $\beta \in (0, 1)$, multiplied by another given positive scalar, such that x^{r+1} given by (4.1) satisfies

$$f(x^r) - f(x^{r+1}) \geq \kappa \langle \nabla f(x^r), x^r - x^{r+1} \rangle,$$

for some other given scalar $\kappa \in (0, 1)$; and (iii) the Goldstein rule (see [Gol74] for details).

The above gradient projection algorithm has been studied very extensively (see, e.g., [Ber76], [BeG82], [CaM87], [Che84], [Dun81], [Dun87], [GaB82], [GaB84], [Gol64],

[Gol74], [LeP65], [McT72]). Typically, a rate of convergence result for this algorithm requires some type of strong convexity assumption on f (see [Dun81], [Dun87], [LeP65]). An exception to this is a result of Bertsekas and Gafni [BeG82] which establishes linear convergence of the iterates under problem assumptions very similar to ours. On the other hand, their result applies only when the stepsizes are small, whereas the linear convergence result below applies for the general stepsize rule (4.2). This is an important distinction since, in general, it is impractical to use small stepsizes. To prove the latter, we show that the iterates satisfy the convergence conditions given in Corollary 3.1.

THEOREM 4.1. *Let $\{x^r\}$ be a sequence generated by the gradient projection algorithm (4.1)–(4.2). Then, $\{x^r\}$ satisfies the hypothesis of Corollary 3.1 and hence converges at least linearly to an element of X^* .*

Proof. From (4.2) and $\alpha^r \leq \bar{\alpha}$ for all r we see that

$$f(x^r) - f(x^{r+1}) \geq \frac{\sigma}{\bar{\alpha}} \|x^r - x^{r+1}\|^2, \quad \forall r,$$

so (3.1) holds with $\kappa_1 = \bar{\alpha}/\sigma$ and $r_0 = 0$.

Next, we show that (3.3) holds. It is known that, for any $x \in X$ and any $d \in \mathbb{R}^n$, the quantity $\|x - [x - \alpha d]^+\|$ is monotonically increasing with $\alpha > 0$ and the quantity $\|x - [x - \alpha d]^+\|/\alpha$ is monotonically decreasing with $\alpha > 0$ (see [Lem. 1, GaB84] or [Lem. 2.2, CaM87]), so that

$$\|x - [x - \alpha d]^+\| \geq \min\{1, \alpha\} \|x - [x - d]^+\|, \quad \forall \alpha > 0.$$

Applying the above bound with $x = x^r$, $d = \nabla f(x^r)$, and then using (4.1) yields

$$\|x^r - x^{r+1}\| = \|x^r - [x^r - \alpha^r \nabla f(x^r)]^+\| \geq \min\{1, \alpha^r\} \|x^r - [x^r - \nabla f(x^r)]^+\|, \quad \forall r.$$

Since $\alpha^r \geq \underline{\alpha}$ for all r , this together with (3.1) implies that (3.3) holds with $\kappa_3 = \kappa_1/\min\{1, \bar{\alpha}^2\}$ and $r_0 = 0$.

Finally, it is easily seen from (4.1) that (3.6) holds with α^r as given and with $e^r = 0$ for all r and $r_1 = 0$. \square

5. Linear convergence of a matrix splitting algorithm. In this section we make (in addition to Assumptions 1.1 and 1.2) the following assumption on f .

ASSUMPTION 5.1. f is a convex quadratic function, i.e.,

$$(5.1) \quad f(x) = \frac{1}{2} \langle x, Mx \rangle + \langle q, x \rangle, \quad \forall x \in \mathbb{R}^n,$$

where M is some $n \times n$ symmetric positive semidefinite matrix, and q is some n -vector.

(Such f is of the form (1.1) because any symmetric positive semidefinite matrix can be expressed as $E^T E$ for some matrix E .) If in addition X is the nonnegative orthant in \mathbb{R}^n , then (1.2) reduces to the well-known symmetric monotone linear complementarity problem.

Let (B, C) be a *regular splitting* of M (see, e.g., [Kel65], [LiP87], [OrR70]), i.e.,

$$(5.2) \quad M = B + C, \quad B - C \text{ is positive definite},$$

and consider the following algorithm for solving this special case of (1.2).

MATRIX SPLITTING ALGORITHM. At the r th iteration we are given an $x^r \in X$ (x^0 is chosen arbitrarily), and we compute a new iterate x^{r+1} in X satisfying

$$(5.3) \quad x^{r+1} = [x^{r+1} - Bx^{r+1} - Cx^r - q + h^r]^+,$$

where h^r is some n -vector. (One simple choice for (B, C) is $B = \mu I$ and $C = M - \mu I$, where μ is a fixed scalar greater than $\|M\|/2$.)

The problem of finding an x^{r+1} satisfying (5.3) may be viewed as an affine variational inequality problem, whereby x^{r+1} is the vector $y \in X$ which satisfies the variational inequality

$$(5.4) \quad \langle z - y, By + Cx^r + q - h^r \rangle \geq 0, \quad \forall z \in X.$$

Because B is positive definite ($2B$ is the sum of a positive definite matrix $B - C$ and a positive semidefinite matrix M), the above variational inequality problem always admits a unique solution (see, e.g., [BeT89], [KiS80]). Thus, the iterates $\{x^r\}$ are well defined.

The vector x^{r+1} may be viewed as an inexact solution of the affine variational inequality problem

$$(5.5) \quad y = [y - By - Cx^r - q]^+,$$

with h^r as the associated “error” vector (so $h^r = 0$ corresponds to an exact solution). The idea of introducing the error vector h^r in this manner is adopted from Mangasarian [Man90]. We remark that in some special cases, such as when X is a box and B is lower triangular, (5.5) can be solved exactly (see [LiP87]), but in general this is not possible. Let γ denote the smallest eigenvalue of the symmetric part of $B - C$ (which by hypothesis is positive) and let ϵ be a fixed scalar in $(0, \gamma/2]$. We will consider the following restriction on h^r governing how fast h^r tends to zero:

$$(5.6) \quad \|h^r\| \leq (\gamma/2 - \epsilon)\|x^r - x^{r+1}\|, \quad \forall r.$$

It can be verified, by using the linear convergence property of $\{x^r\}$ shown below (see Theorem 5.1), that (5.6) is a special case of the criteria introduced by Mangasarian for the case of a symmetric splitting (see [Man90, Alg. 2.1]). This criterion can be met, for example, by appropriately terminating any iterative method used to solve (5.5). To illustrate, fix r and suppose that we have a sequence of points converging to the solution of (5.5). (Methods for generating such a sequence are described in, for example, [BeT89].) Suppose that the limit is not x^r (otherwise x^r is already in X^*) and let $F : \mathbb{R}^n \mapsto \mathbb{R}^n$ be the continuous function given by $F(y) = [y - By - Cx^r - q]^+$. Then, for all points y sufficiently far along in this sequence we have

$$\|(I - B)(y - F(y))\| \leq (\gamma/2 - \epsilon)\|x^r - F(y)\|.$$

(This is because the limit, say \bar{y} , is not equal to x^r and satisfies $\bar{y} = F(\bar{y})$.) Take any such point y and set

$$h^r = (I - B)(y - F(y)), \quad x^{r+1} = F(y).$$

Then, h^r and x^{r+1} satisfy (5.3) and (5.6).

In the special case where X is a box, the above matrix splitting algorithm has been very well studied (see [Man77], [Pan82], [Pan84], [Pan86], [LuT89a]). But, even in this case, very little is known about its rate of convergence. Only very recently was it shown that the iterates generated by this algorithm indeed converge (see [LuT89a]). If the matrix splitting corresponds to an SOR iteration, then it is known that the sequence $\{Mx^r\}$ converges at least linearly (see [IuD90]). Below we improve on the above results by showing that the iterates converge at least linearly. Moreover, our result holds for the general case where X is any polyhedral set, not just a box. We remark that, from a theoretical standpoint, the general polyhedral case is no harder to treat (using our analysis) than the box case. However, from a practical standpoint, the box case is typically easier to work with.

THEOREM 5.1. *Let $\{x^r\}$ be the sequence generated by the matrix splitting algorithm (5.2)–(5.3), (5.6). Then $\{x^r\}$ satisfies the hypothesis of Corollary 3.1 and hence converges at least linearly to an element of X^* .*

Proof. We first verify that (3.1) holds. Fix any r . Since x^{r+1} satisfies the variational inequality (5.4), then, by plugging in x^r for z and x^{r+1} for y in (5.4), we obtain

$$\langle x^{r+1} - x^r, Bx^{r+1} + Cx^r + q - h^r \rangle \leq 0.$$

Also, from $M = B + C$ (cf. (5.2)) and our choice of f (cf. (5.1)) we have that

$$f(x^{r+1}) - f(x^r) = \langle x^{r+1} - x^r, Bx^{r+1} + Cx^r + q \rangle + \langle x^{r+1} - x^r, (C - B)(x^{r+1} - x^r) \rangle / 2.$$

Combining the above two relations then gives

$$\begin{aligned} f(x^{r+1}) - f(x^r) &\leq \langle x^{r+1} - x^r, h^r \rangle + \langle x^{r+1} - x^r, (C - B)(x^{r+1} - x^r) \rangle / 2 \\ &\leq \|x^{r+1} - x^r\| \|h^r\| - \gamma \|x^{r+1} - x^r\|^2 / 2 \\ &\leq -\epsilon \|x^{r+1} - x^r\|^2, \end{aligned}$$

where the last inequality follows from (5.6). Hence (3.1) holds with $\kappa_1 = 1/\epsilon$ and $r_0 = 0$.

We now show that (3.3) holds. From (5.3) we have that

$$\begin{aligned} \|x^r - [x^r - \nabla f(x^r)]^+\| &= \|x^r - [x^r - Mx^r - q]^+\| \\ &= \|x^r - [x^r - Mx^r - q]^+ - x^{r+1}\| \\ &\quad + \|x^{r+1} - Bx^{r+1} - Cx^r - q + h^r\|^+ \\ &\leq \|x^r - x^{r+1}\| + \|[x^r - Mx^r - q]^+ \\ &\quad - [x^{r+1} - Bx^{r+1} - Cx^r - q + h^r]^+\| \\ &\leq 2\|x^r - x^{r+1}\| + \|Mx^r - Bx^{r+1} - Cx^r + h^r\| \\ &\leq 2\|x^r - x^{r+1}\| + \|B(x^r - x^{r+1})\| + \|h^r\| \\ &\leq (2 + \|B\| + \gamma/2)\|x^r - x^{r+1}\|, \end{aligned}$$

where the second inequality follows from the nonexpansive property of the projection operator $[\cdot]^+$, the third inequality follows from $M = B + C$, and the last inequality follows from (5.6). This together with (3.1) shows that (3.3) holds with $\kappa_3 = (2 + \|B\| + \gamma/2)^2 \kappa_1$ and $r_0 = 0$.

Finally, we show that (3.6) holds with $\alpha^r = 1$ for all r and some sequence of n -vectors $e^r \rightarrow 0$. From (5.3), $\nabla f(x^r) = Mx^r + q$ (cf. (5.1)) and $M = B + C$ (cf. (5.2)) we have

$$x^{r+1} = [x^r - \nabla f(x^r) + B(x^r - x^{r+1}) + h^r]^+, \quad \forall r,$$

so (3.6) holds with $\alpha^r = 1$ and $e^r = B(x^r - x^{r+1}) + h^r$ for all r . Since f is bounded from below on X (cf. Assumption 1.2 (a)), then (3.1) implies $x^r - x^{r+1} \rightarrow 0$. Hence $h^r \rightarrow 0$ (cf. (5.6)) and therefore $e^r \rightarrow 0$. \square

Note that we can allow the matrix splitting (B, C) to vary from iteration to iteration, provided that the eigenvalues of the symmetric part of $B - C$ are bounded from above and are bounded away from zero.

6. Conclusion and extensions. In this paper we have presented a general framework for establishing the linear convergence of descent methods for solving (1.2) and have applied it to two well-known algorithms: the gradient projection algorithm of Goldstein and Levitin and Polyak, and the matrix splitting algorithm using regular splitting. The key to this framework lies in a new bound for estimating the distance from a feasible point to the optimal solution set.

There are a number of directions in which our results can be extended. For example, the bound of Theorem 2.1 can be shown to hold locally for any quadratic (possibly nonconvex) function f , which enables us to extend the results of §5 to symmetric non-monotone linear complementarity problems. In fact, the same bound can be shown to hold locally for nonsymmetric linear complementarity problems as well. (An example given by Mangasarian and Shiau [MaS86, Ex. 2.10] shows that this bound does not hold globally for nonsymmetric problems.) We hope to report on these extensions in the future. Finally, it would be worthwhile to find descent methods, other than those treated here, to which our linear convergence framework can be fruitfully applied. In fact, we have been able to apply this framework to the coordinate descent algorithm to obtain a proof of its linear convergence simpler than that given in [LuT89b].

Acknowledgment. We gratefully acknowledge the comments of the referees which led to significant improvements in the presentation of this paper.

REFERENCES

- [Ber76] D. P. BERTSEKAS, *On the Goldstein-Levitin-Polyak gradient projection method*, IEEE Trans. Automat. Contr., AC-21 (1976), pp. 174-184.
- [BeG82] D. P. BERTSEKAS AND E. M. GAFNI, *Projection methods for variational inequalities with application to the traffic assignment problem*, Math. Prog. Study, 17 (1982), pp. 139-159.
- [BeT89] D. P. BERTSEKAS AND J. N. TSITSIKLIS, *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
- [CaM87] P. H. CALAMAI AND J. J. MORÉ, *Projected gradient methods for linearly constrained problems*, Math. Programming, 39 (1987), pp. 93-116.
- [Che84] Y. C. CHENG, *On the gradient-projection method for solving the nonsymmetric linear complementarity problem*, J. Optim. Theory Appl., 43 (1984), pp. 527-541.
- [Cen88] Y. CENSOR, *Parallel application of block-iterative methods in medical imaging and radiation therapy*, Math. Programming Ser. B, 42 (1988), pp. 307-325.
- [CeL87] Y. CENSOR AND A. LENT, *Optimization of "log x" entropy over linear equality constraints*, SIAM J. Control Optim., 25 (1987), pp. 921-933.
- [CoP82] R. W. COTTLE AND J.-S. PANG, *On the convergence of a block successive over-relaxation method for a class of linear complementarity problems*, Math. Prog. Study, 17 (1982), pp. 126-138.
- [Cry71] C. W. CRYER, *The solution of a quadratic programming problem using systematic overrelaxation*, SIAM J. Control Optim., 9 (1971), pp. 385-392.
- [Dun81] J. C. DUNN, *Global and asymptotic convergence rate estimates for a class of projected gradient processes*, SIAM J. Control Optim., 19 (1981), pp. 368-400.
- [Dun87] ———, *On the convergence of projected gradient processes to singular critical points*, J. Optim. Theory Appl., 55 (1987), pp. 203-216.

- [FiM68] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley and Sons, New York, 1968.
- [Fri75] B. R. FRIEDEN, *Image enhancement and restoration*, in *Picture Processing and Digital Filtering*, T. S. Huang, ed., Springer-Verlag, New York, Berlin, Heidelberg, Tokyo, (1975), pp. 177–248.
- [GaB82] E. M. GAFNI AND D. P. BERTSEKAS, *Convergence of a gradient projection method*, Laboratory for Information and Decision Systems Report No. P-1201, Massachusetts Institute of Technology, Cambridge, MA, 1982.
- [GaB84] ———, *Two-metric projection methods for constrained optimization*, SIAM J. Control Optim., 22 (1984), pp. 936–964.
- [GMSTW86] P. E. GILL, W. M. MURRAY, M. A. SAUNDERS, J. A. TOMLIN, AND M. A. WRIGHT, *On projected newton barrier methods for linear programming and an equivalence to Karmarkar's projective methods*, Math. Programming, 36 (1986), pp. 183–209.
- [Gol64] A. A. GOLDSTEIN, *Convex programming in Hilbert space*, Bull. Am. Math. Soc., 70 (1964), pp. 709–710.
- [Gol74] ———, *On gradient projection*, Proc. 12th Ann. Allerton Conference on Circuits and Systems, Allerton Park, IL, (1974), pp. 38–40.
- [Her80] G. T. HERMAN, *Image Reconstruction from Projection: The Fundamentals of Computerized Tomography*, Academic Press, New York, 1980.
- [Hil57] C. HILDRETH, *A quadratic programming procedure*, Naval Res. Logist., 4 (1957), pp. 79–85; see also *Erratum*, Naval Res. Logist., 4 (1957), p. 361.
- [Hof52] A. J. HOFFMAN, *On approximate solutions of systems of linear inequalities*, J. Res. Natl. Bur. Standards, 49 (1952), pp. 263–265.
- [IuD90] A. N. IUSEM AND A. DE PIERRO, *On the convergence properties of Hildreth's quadratic programming algorithm*, Math. Programming, 47 (1990), pp. 37–51.
- [Jay82] E. T. JAYNES, *On the rationale of maximum entropy methods*, Proc. IEEE, 70 (1982), pp. 939–952.
- [JoS84] R. JOHNSON AND J. E. SHORE, *Which is the better entropy expression for speech processing: $-S \log S$ or $\log S$?*, IEEE Trans. Acoust. Speech Sign. Proc., ASSP-32 (1984), pp. 129–136.
- [Kel65] H. B. KELLER, *On the solution of singular and semidefinite linear systems by iteration*, SIAM J. Numer. Anal., 2 (1965), pp. 281–290.
- [KiS80] D. KINDERLEHRER AND G. STAMPACCHIA, *An Introduction to Variational Inequalities and Applications*, Academic Press, New York, 1980.
- [Kru37] J. KRUTHOF, *Calculation of Telephone Traffic*, De Ingenieur (Elektrotechnik 3), 52 (1937), pp. E15–25.
- [LaS81] B. LAMOND AND N. F. STEWART, *Bregman's balancing method*, Transportation Res. Part-B, 15B (1981), pp. 239–248.
- [LeP65] E. S. LEVITIN AND B. T. POLYAK, *Constrained minimization methods*, Z. Vycisl. Mat. i Mat. Fiz., 6 (1965), pp. 787–823. English translation in USSR Comput. Math. and Math. Phys., 6 (1965), pp. 1–50.
- [LiP87] Y. Y. LIN AND J.-S. PANG, *Iterative methods for large convex quadratic programs: A survey*, SIAM J. Control Optim., 25 (1987), pp. 383–411.
- [LuT89a] Z.-Q. LUO AND P. TSENG, *On the convergence of a matrix-splitting algorithm for the symmetric monotone linear complementarity problem*, SIAM J. Control Optim., 29 (1991), pp. 1037–1060.
- [LuT89b] ———, *On the convergence of the coordinate descent method for convex differentiable minimization*, Laboratory for Information and Decision Systems Report No. P-1924, Massachusetts Institute of Technology, Cambridge, MA (1989; revised 1990), J. Optim. Theory Appl., 72 (1992), pp. 7–35.
- [Man77] O. L. MANGASARIAN, *Solution of symmetric linear complementarity problems by iterative methods*, J. Optim. Theory Appl., 22 (1977), pp. 465–485.
- [Man88] ———, *A simple characterization of solution sets of convex programs*, Oper. Res. Lett., 7 (1988), pp. 21–26.
- [Man90] ———, *Convergence of iterates of an inexact matrix splitting algorithm for the symmetric monotone linear complementarity problem*, Computer Sciences Technical Rep. #917, University of Wisconsin, Madison, WI (1990), SIAM J. Optimization, 1 (1991), pp. 114–122.

- [MaD87] O. L. MANGASARIAN AND R. DE LEONE, *Parallel successive overrelaxation methods for symmetric linear complementarity problems and linear programs*, J. Optim. Theory Appl., 54 (1987), pp. 437–446.
- [MaD88a] ———, *Parallel gradient projection successive overrelaxation for symmetric linear complementarity problems and linear programs*, Ann. Oper. Res., 14 (1988), pp. 41–59.
- [MaD88b] ———, *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 (1988), pp. 1–14.
- [MaS86] O. L. MANGASARIAN AND T.-H. SHIAU, *Error bounds for monotone linear complementarity problems*, Math. Programming, 36 (1986), pp. 81–89.
- [MaS87] ———, *Lipschitz continuity of solutions of linear inequalities, programs and complementarity problems*, SIAM J. Control Optim. 25 (1987), pp. 583–595.
- [McT72] G. P. MCCORMICK AND R. A. TAPIA, *The gradient projection method under mild differentiability conditions*, SIAM J. Control Optim., 10 (1972), pp. 93–98.
- [Mor89] J. J. MORÉ, *Gradient projection techniques for large-scale optimization problems*, Proc. of the 28th Conf. on Decision and Control, Tampa, Florida (December 1989).
- [OrR70] J. M. ORTEGA AND W. C. RHEINOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, (1970).
- [Pan82] J.-S. PANG, *On the convergence of a basic iterative method for the implicit complementarity problem*, J. Optim. Theory Appl., 37 (1982), pp. 149–162.
- [Pan84] ———, *Necessary and sufficient conditions for the convergence of iterative methods for the linear complementarity problem*, J. Optim. Theory Appl., 42 (1984), pp. 1–17.
- [Pan86] ———, *More results on the convergence of iterative methods for the symmetric linear complementarity problem*, J. Optim. Theory Appl., 49 (1986), pp. 107–134.
- [Pan87] ———, *A posteriori error bounds for the linearly-constrained variational inequality problem*, Math. Oper. Res., 12 (1987), pp. 474–484.
- [Pow88] M. J. D. POWELL, *An algorithm for maximizing entropy subject to simple bounds*, Math. Programming, 42 (1988), pp. 171–180.
- [Rob73] S. M. ROBINSON, *Bounds for error in the solution set of a perturbed linear program*, Linear Algebra Appl., 6 (1973), pp. 69–81.
- [Roc70] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
- [Son88] G. SONNEVEND, *New algorithms in convex programming based on a notion of center (for systems of analytic inequalities) and on rational extrapolation*, in Trends in Mathematical Optimization, K.-H. Hoffman, J.-B. Hiriart-Urruty, J. Zowe, and C. Lemarechal, eds., Birkhäuser-Verlag, Basel, Switzerland, 1988, pp. 311–326.
- [Tse89] P. TSENG, *Descent methods for convex essentially smooth minimization*, Laboratory for Information and Decision Systems Report No. P-1896, Massachusetts Institute of Technology, Cambridge, MA (1989; revised 1990); J. Optim. Theory Appl., 71 (1991), pp. 425–463.
- [Tse90] ———, *Dual ascent methods for problems with strictly convex costs and linear constraints: A unified approach*, SIAM J. Control Optim., 28 (1990), pp. 214–242.
- [TsB87a] P. TSENG AND D. P. BERTSEKAS, *Relaxation methods for problems with strictly convex separable costs and linear constraints*, Math. Programming, 38 (1987), pp. 303–321.
- [TsB87b] ———, *Relaxation methods for problems with strictly convex costs and linear constraints*, Laboratory for Information and Decision Systems Report No. P-1717, Massachusetts Institute of Technology, Cambridge, MA (1987); Math. Oper. Res., 16 (1991), pp. 462–481.

ON THE OPTIMAL TRACKING PROBLEM*

OFER ZEITOUNI ^{†‡} AND MOSHE ZAKAI ^{†§}

Abstract. The problem of optimally tracking a stochastic process based on noisy measurements through a window is being considered. Such a problem arises in a variety of applications where the field of view of the measuring device is limited (e.g., the “aiming” problem). With the objective to minimize the probability to lose track, the problem is formulated as an optimal control problem with partial observations. The existence of an optimal control is proven for both cases of discrete and continuous time (diffusion) signal process observed in white noise. The low observation noise asymptotics are then considered: for a one-dimensional problem, a proposed suboptimal tracker is shown to be asymptotically logarithmically optimal.

Key words. stochastic control, nonlinear filtering, tracking

AMS(MOS) subject classifications. 93E20, 93E11

1. Introduction. Consider a typical tracking system as described in Fig. 1. Here, x_t denotes the (angular) location of a target. A telescope is “attempting” to track the target by changing its boresight angle u_t . Due to the (unknown) motion of the target, and possibly that of the platform on which the telescope is located, the angular position of the target is assumed to be random. The information available for tracking is the image of the target on a light-sensitive screen, whose output depends on the angle between the boresight and target directions. A typical error sensitivity curve $h(x - u)$ is shown in Fig. 1. This signal is then used to rotate the telescope towards the target. Due to the limited field of view of the telescope, information is lost if the pointing error $|x_t - u_t|$ exceeds a level c . The input available to the controller is a noisy version \dot{y}_t of the error signal, namely $\dot{y}_t = h(x_t - u_t) + \dot{v}_t$ where \dot{v}_t is the observation noise, which is generated by the photocell and is independent of the x_t process.

We can model the above problem as follows. Let $x_t \in R^k$ satisfy the stochastic Ito differential equation

$$(1.1) \quad dx_t = m(x_t) dt + \sigma(x_t) dw_t, \quad x_0 = x.$$

The observation $y_t \in R$ satisfies

$$(1.2) \quad dy_t = h(x_t - u_t) dt + \epsilon dv_t, \quad y_0 = 0,$$

where w , v are independent standard Brownian motions, and $\epsilon > 0$ is a constant. The control u_t has to be a functional of t and of the path $\{y_\theta, 0 \leq \theta \leq t\}$, which will

*Received by the editors April 18, 1990; accepted for publication (in revised form) March 8, 1991.

[†]Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa 32000, Israel.

[‡]Part of the work of this author was done while the author visited the Laboratory for Information and Decision Systems, MIT, under support from US Army Contract DAAL03-86-K-0171.

[§]The work of this author was partially supported by the fund for the promotion of research at the Technion.

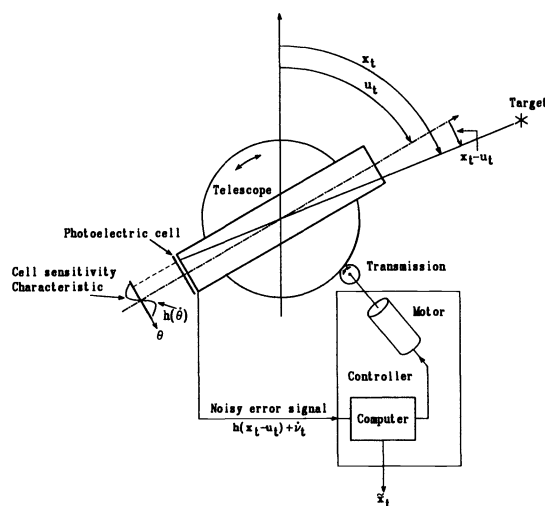


FIG. 1

be denoted y_o^t , namely $u_t = u(t, y_o^t)$. A slightly more general control strategy will be considered later on.

The control functional $u(t, y_o^t)$ has to be chosen so as to achieve the following goals:

1. To keep the target in the field of view for as long as possible.
2. To supply a reliable estimate \hat{x}_t of the target's position x_t .
3. To detect a loss of lock.

In this paper we only deal with the first goal, addressing it as an optimization problem. However, the results in §3 indicate that, in the case of high signal-to-noise ratios (ϵ small), an asymptotically optimal control for goal 1 above becomes (under appropriate assumptions) an asymptotically optimal control for goal 2.

We note that in the model of (1.1), (1.2), we have not taken into account any dynamics limitations on the control u . Such limitations appear, for example, when the inertia of the tracking antenna is taken into account. These limitations, while tractable, complicate the analysis and will not be considered in this paper. We observe, however, that in the case of a phase or delay locked loop, such physical limitations do not appear.

The tracking model described above is related not only to the optical tracker of Fig. 1 but also to other systems of technical significance, such as automatic range and angle trackers in radar systems, phase locked and delay locked loops in communication systems, etc. For a guide to applications of tracking systems, c.f. [1].

The design of systems of the type discussed in this paper is usually based on a linearization of $h(x)$ for all x and the application of extended Kalman filtering or Linear Quadratic Gaussian (LQG) techniques. This approach yields, no doubt, very good trackers. However, in general, the linearization approach does not seem to minimize the probability to lose track. In this paper, we focus on this latter minimization. Other tracking models have been analyzed in the literature quite extensively (cf. [1]), however it seems that this optimization framework has not been previously considered. Some recent papers deal with the related (though different) problem of controlling the escape probability of dynamical systems ([6], [11], [12] for the linear case). These

papers deal with the asymptotic ($\epsilon \rightarrow 0$) problem of controlling the system

$$dx^{u,\epsilon} = f(x^{u,\epsilon}, u)dt + \epsilon \sigma(x^{u,\epsilon})dw_t$$

so as to keep $x^{u,\epsilon}$ in a given set G for as long as possible. The tracking problem, even in its asymptotics, differs fundamentally from this model in that the process noise is not assumed to be small.

The existence of an optimal stochastic control for the model of (1.1), (1.2) seems at first sight to be a direct consequence of known results for the partially observed control problem. However, due to the presence of a control in the observation, the model does not conform with known results unless appropriate modifications are introduced. These are carried out in §2, where the optimal control aspects of the tracking problem are discussed. Following Fleming and Pardoux [9], we define the notion of weak sense controllers. We then modify the results of [9] and [4] to show the existence of an optimal weak controller in the case where $h(\cdot)$ is a linear function in the interior of its support. The Appendix discusses the optimal stochastic control problem defined by a discrete time model analogous to the one described above. Existence of solutions, the characterization of the optimizer, as well as some suboptimal controllers are presented.

In §3, which contains the main results of this paper, we derive upper and lower bounds on the performance for the continuous time case. In the particular case where

$$h(\theta) = \begin{cases} h_o \cdot \theta, & |\theta| \leq c, \\ 0, & \text{otherwise,} \end{cases}$$

the upper and lower bounds are asymptotically ($\epsilon \rightarrow 0$) logarithmically tight, leading to explicit asymptotically optimal controls. These bounds are based on appropriately defined discrete time versions of the continuous time models, on bounds on the nonlinear filtering problem, and on large deviations arguments. It is clear that an analogous version of the results could be derived using the same methods for the discrete time case.

Section 3 can be understood without following the proofs in §2.

2. The optimal control aspects of the tracking problem. We assume throughout that the coefficients in (1.1), (1.2) $m(\cdot)$, $\sigma(\cdot)$ are C_b^2 functions of appropriate dimensions, w_t , v_t are standard independent \mathcal{F}_t Brownian motions of dimensions $k, 1$, respectively, $h(\cdot)$ is a continuously differentiable function of compact support $2c$ (i.e., $h(z) \equiv 0$ for $|z| > c$), and $u \in U$, a convex set in \mathbb{R}^k that needs not necessarily be compact. The performance criterion that we want to maximize is

$$(2.1) \quad J_T^u = E(1_{\|x-u\|_T \leq c}).$$

Following [9], we adopt the following formulation: Let $\Omega_1 = C[[0, T]; \mathbb{R}^k]$, $\Omega_2 = C_o[[0, T]; \mathbb{R}]$, and $\Omega_3 = L^\infty([0, T]; \mathbb{R}^k)$. We endow $\Omega_1 \times \Omega_2$ with the supremum norm topology and Ω_3 with the weak topology, and denote $\Omega = \Omega_1 \times \Omega_2 \times \Omega_3$. The canonical element of Ω is denoted (x_t, y_t, u_t) .

Under our assumptions, there exists a strong solution to (1.1), and for each pair of paths $y \in \Omega_2$, $u \in \Omega_3$, there exists a regular conditional probability distribution (rcpd) $P_x^{y,u}$ over $(\Omega_1, \mathcal{B}(\Omega_1))$. Finally, define $\tilde{v}_t = \int_0^t u_s ds$ and let

$$\mathcal{F}_t^1 = \mathcal{F}_t(w), \quad \mathcal{F}_t^2 = \mathcal{F}_t(y) \times \mathcal{F}_t(\tilde{v}).$$

DEFINITION. An admissible wide sense control π is a probability measure on $(\Omega_2 \times \Omega_3, \mathcal{F}_T^2)$ which makes y into an \mathcal{F}_t^2 Brownian motion. The set of admissible controls is denoted Π . The resulting measure on Ω is denoted $\overset{\circ}{P}_\pi(x, y, u)$.

Next define

$$(2.2) \quad z_t = \exp \left(\int_0^t h(x_s - u_s) dy_s - \frac{1}{2} \int_0^t h^2(x_s - u_s) ds \right).$$

Since h is bounded, z_t is an L^1 martingale, and therefore, one may define a new probability measure P_π by

$$(2.3) \quad \frac{dP_\pi(x, y, u)}{d\overset{\circ}{P}_\pi(x, y, u)} = z_T.$$

Note that if $d\pi(y, u) = d\pi(y)\delta(u - g(y))$, where $g(y)$ is a progressively measurable functional on $C[[0, T]; \mathbb{R}]$, then (x, y, u) solves, under P_π , (1.1)–(1.2). In such cases, we will say that u is a *strict sense control*, and the set of such π' is denoted Π^s . Finally, the criterion J^u can be written alternatively as

$$(2.4) \quad J_\pi = E_\pi(1_{\|x-u\|_T \leq c}) = \overset{\circ}{E}_\pi(z_T 1_{\|x-u\|_T \leq c}),$$

where the optimization goal on hand is to compute

$$J^u \triangleq \sup_{\pi \in \Pi} J_\pi.$$

We note that, by an adaptation of the results of [5],

$$(2.5) \quad J^u = \sup_{\pi \in \Pi} J_\pi = \sup_{\pi \in \Pi^s} J_\pi,$$

and moreover, a “separation principle” [9] holds. However, the model (1.1)–(1.2) does not conform directly with the standard partially observed control literature [9], [4], [2], due to the existence of controls in the observation (1.2). To obtain results on the existence of optimal controls, we will restrict the observation function, as follows. Assume that

$$(2.6) \quad h(x) = \begin{cases} h^T x, & \|x\| \leq c, \\ 0, & \|x\| > c \end{cases}$$

The restriction to truncated linear observation functions as in (2.6) will be discussed in the remark following corollary 2.1. We now prove Lemma 2.1.

LEMMA 2.1. *The maximization problem (2.5) is equivalent to the problem of maximizing*

$$\tilde{J}^u = \sup_{\pi \in \Pi} \tilde{J}_\pi$$

where

$$(2.7) \quad \tilde{J}_\pi = \overset{\circ}{E}_\pi(\tilde{z}_T 1_{\|x-u\|_T \leq c})$$

and

$$(2.8) \quad \tilde{z}_t = \exp \left(\int_0^t h^T x_s dy_s - \frac{1}{2} \int_0^t |h^T x_s|^2 ds \right).$$

Proof. Let π^n be a maximizing sequence for (2.5). Define a new sequence of measures by the change of variables $y \rightarrow y + \int_0^\cdot u_s ds$, i.e.,

$$d\tilde{\pi}_n(y, u) = d\pi_n(y + v, u).$$

Clearly, $\tilde{J}_{\pi^n} = J_{\pi^n}$, and therefore $J^u \leq \tilde{J}^u$. The converse is obtained similarly and the lemma is established. \square

Remark. The same lemma holds for strict sense controls, i.e.,

$$\sup_{\pi \in \Pi^s} J_\pi = \sup_{\pi \in \Pi^s} \tilde{J}_\pi = J^u.$$

Lemma 2.1 enables us to apply the results of [4] to obtain Theorem 2.1.

THEOREM 2.1. *Assume U is compact. Then an optimal wide sense control π exists.*

Proof. In light of [4, Thm. 2.11] and of Lemma 2.1, we have only to check that $1_{\|x-u\|>c}$ is, for each $x_t \in \Omega_1$, lower semicontinuous with respect to the weak topology on L_∞ . However, $u_n \xrightarrow{w} u$ implies that $\liminf \|x - u_n\| \geq \|x - u\|$ and therefore $\liminf 1_{\|x-u_n\|>c} \geq 1_{\|x-u\|>c}$, and the conclusion follows. \square

We obtain the following corollary as well.

COROLLARY 2.1. *Theorem 2.1 continues to hold when $U = \mathbb{R}^k$.*

Proof. It suffices to show that the set of probability measures

$$\pi^* = \{\pi \mid J_\pi^u \leq J^u + 1\}$$

is tight. This follows from the fact that

$$P_x(\|x\| > n) \xrightarrow{n \rightarrow \infty} 0 \text{ and } P_v(\|v\| > n) \xrightarrow{n \rightarrow \infty} 0. \quad \square$$

Remark. We comment here on the particular choice of $h(\cdot)$ in (2.6). To simplify matters, consider the case $k = 1$. The proof of the lower semicontinuity of Theorem 3.1 in [9], [2], [4] relied heavily on the “pathwise transformation”

$$\int_0^t h(x_s) dy_s = y_t h(x_t) - \int_0^t y_s h'(x_s) dx_s - \frac{1}{2} \int_0^t y_s h''(x_s) \sigma^2(x_s) ds,$$

which had been used to obtain continuity results of the conditional density. In general, due to the u_t term in z_T , one cannot use this transformation without structural restrictions on u_t , which we prefer not to assume in this stage. Moreover, under the weak topology on u , the convergence of $u_n \xrightarrow{w} u$ does not imply convergence of $\int_0^t h(x_t - u_n(t)) dt$ to $\int_0^t h(x_t - u_t) dt$ or even semicontinuity in any direction. This leads to the (practically important) case (2.6). Note that under more stringent assumptions on u , e.g., $u \in C^1[0, T]$, one may relax that assumption. However, one loses then the compactness properties and consequently the existence theorem.

We conclude this section by pointing out that, for sufficiently smooth controls (and in particular C^1 controls), a separated control solution of the problem can be constructed. The interested reader is referred to [20].

3. High signal-to-noise ratio (SNR) asymptotics. In this section we consider the one-dimensional ($k = 1$) version of the problem (1.1), (1.2) with $\sigma(\cdot) \equiv 1$ and a truncated linear observation function. For the high SNR case, i.e., $\epsilon \rightarrow 0$, we propose a suboptimal filter that is based on the approximation to the conditional mean and show, by bounding techniques, that its performance in a suitable logarithmic sense, is asymptotically optimal. Specifically, we show that, for *any* admissible control rule u and any $T > 0$,

$$(3.1) \quad \liminf_{\epsilon \rightarrow 0} \epsilon \log \inf P^u(\|x - u\|_T > c) \geq -\frac{c^2 h}{2},$$

whereas, using the suboptimal filter \hat{u} described below,

$$(3.2) \quad \lim_{\epsilon \rightarrow 0} \epsilon \log P^{\hat{u}}(\|x - \hat{u}\|_T > c) = -\frac{c^2 h}{2},$$

from which the asymptotic log optimality of \hat{u} follows. Moreover, letting τ^ϵ denote the first exit time $\tau^\epsilon \triangleq \{\inf t : |x_t - u_t| > c\}$, it holds that for any admissible control law,

$$(3.1a) \quad \lim_{\epsilon \rightarrow 0} \epsilon \log E\tau^\epsilon \leq \frac{c^2 h}{2},$$

and for the particular control \hat{u} ,

$$(3.2a) \quad \lim_{\epsilon \rightarrow 0} \epsilon \log E\tau^\epsilon = \frac{c^2 h}{2}.$$

In the sequel, we rewrite the model and prove the asymptotic results. A discussion of the significance of various parameters in the design follows.

Consider the state equation where $\sigma(\cdot) \equiv 1$ and $m(\cdot)$ is a C_b^3 function,

$$(3.3) \quad dx_t = m(x_t) dt + dw_t, \quad x_t \in \mathbb{R}^1,$$

and the observation

$$(3.4) \quad dy_t = h \cdot (x_t - u_t) dt + \epsilon dv_t, \quad y_0 = 0.$$

Note that $h(\cdot)$ is a linear gain throughout this section. We have the following two lemmas.

LEMMA 3.1. *Let $u \in \Pi$. Then (3.1) and (3.1a) hold. Moreover, a nonasymptotic lower bound on the performance is given by (3.13).*

LEMMA 3.2. *Let \hat{u}_t satisfy*

$$d\hat{u}_t = m(\hat{u}_t) dt + \frac{1}{\epsilon} dy_t.$$

Then (3.2) and (3.2a) hold.

Proof of Lemma 3.1. Let $n(\epsilon)$ be a given integer, to be determined below, and define $\Delta t(\epsilon) = T/n(\epsilon)$. (In fact, we will prove more than needed; the less patient reader may assume $n(\epsilon) = 1$ throughout the proof.) Let $\tau_i \triangleq i \cdot \Delta t(\epsilon)$, $i = 0, 1, \dots, n(\epsilon)$. Clearly, for any $u \in \Pi$,

$$E(1_{\|x-u\|_T < c}) \leq P\left(\bigcap_i \{|x_{\tau_i} - u_{\tau_i}| < c\}\right).$$

Therefore (note that the classes of controls on both sides of (3.5) below are identical),

$$(3.5) \quad J_\pi = \sup_{u \in \Pi} E(1_{\|x-u\|_T < c}) \leq \sup_{u \in \Pi} P\left(\bigcap_i \{|x_{\tau_i} - u_{\tau_i}| < c\}\right).$$

Let Π^{dt} denote the controls u , which are adapted to

$$\sigma(y_0^t) \vee \mathcal{F}_{t-\Delta t(\epsilon)}^* \quad \text{where } \mathcal{F}_{t-\Delta t(\epsilon)}^* = \sigma(x_s, \quad 0 \leq s \leq t - \Delta t(\epsilon)),$$

namely, the controller “knows” not only y_0^t but also $x_0^{t-\Delta t(\epsilon)}$. From (3.5) and the definition of Π^{dt} , we have

$$(3.6) \quad \begin{aligned} J_\pi &\leq \sup_{u \in \Pi^{dt}} P\left(\bigcap_i \{|x_{\tau_i} - u_{\tau_i}| < c\}\right) \\ &\leq \prod_{i=0}^{n(\epsilon)} \sup_{u_i \in \sigma(y_{\tau_{i-1}}^{\tau_i}) \vee \mathcal{F}_{\tau_{i-1}}^*} P(|x_{\tau_i} - u_{\tau_i}| < c | x_{\tau_{i-1}}) \\ &= \prod_{i=0}^{n(\epsilon)} E\left\{ \sup_{u_i \in \sigma(y_{\tau_{i-1}}^{\tau_i}) \vee \mathcal{F}_{\tau_{i-1}}^*} P(|x_{\tau_i} - u_{\tau_i}| < c | y_{\tau_{i-1}}^{\tau_i}, x_{\tau_{i-1}}) \right\}, \end{aligned}$$

where we have used in (3.6) the Markov structure of (3.3) and that of $\sigma(y_{\tau_{i-1}}^{\tau_i}) \vee \sigma(x_{\tau_{i-1}})$. Since the right-hand side of (3.6) involves only a one step predictor, and the maximization is unaffected by normalization constants, its optimal solution is easily seen to be

$$(3.7) \quad u_i = \arg \max_u \int_{u-c}^{u+c} \rho_{x, \tau_i}^i(z) dz,$$

where $\rho_{x,t}^i(z)$ satisfies the equation ([17]):

$$(3.8) \quad \begin{aligned} d\rho_{x,t}^i(z) &= \frac{1}{2} \frac{\partial^2}{\partial z^2} \rho_{x,t}^i(z) dt - \frac{\partial(m(z)\rho_{x,t}^i(z))}{\partial z} + \frac{1}{\epsilon^2} h z \rho_{x,t}^i(z) dy_t \\ \rho_{x, \tau_{i-1}}^i(z) &= \delta_{x_{\tau_{i-1}}}(z). \end{aligned}$$

By [18, Thms. 1–3], it holds under our assumptions that

$$(3.9) \quad k_1(t) \mathcal{N}(\tilde{x}_i^L(t), \tilde{\sigma}_i^L(t)) \leq \rho_{x,t}^i(z) \leq k_2(t) \mathcal{N}(\tilde{x}_i^U(t), \tilde{\sigma}_i^U(t)),$$

where $\mathcal{N}(a, b)$ denotes a normal distribution, mean a , variance b , $\tilde{x}_i^L(t)$, $\tilde{x}_i^U(t)$ are the output of $y_{\tau_{i-1}}^{\tau_i}$ driven filters such that $|\tilde{x}_i^L(t) - \tilde{x}_i^U(t)|/(\sigma_i^L(t))^{1/2} \rightarrow 0$ as $\epsilon \rightarrow 0$, $\tilde{\sigma}_i^L(t)$, $\tilde{\sigma}_i^U(t)$, deterministic variances, are of the form

$$(3.10) \quad \tilde{\sigma}_i^{U,L}(t) = \frac{\epsilon}{h} \operatorname{tgh}\left(\frac{ht}{\epsilon}\right) + o(\epsilon),$$

(cf. [18, (3.3a)]), and k_1, k_2 satisfy appropriate stochastic equations such that for all $t > 0$, $k_1(t)/k_2(t) \xrightarrow{\epsilon \rightarrow 0} k(t)$, almost surely and in $L^2(\Omega)$, with $k(t) \neq 0$. Note that the results in [18] require a specific initial density parameterized by a variance γ . However, by the same derivation, these results hold for $\gamma = 0$ with the obvious modifications (i.e., using instead of (2.13a) in [18] the equation $d\Lambda_{it}/dt = I - (P_{it} + g^2 N_0^{-1})\Lambda_{it}^2$). Hence, they hold true for the case of Dirac initial condition, which is the case considered in (3.8). Substituting in the above $t = \tau_i$, we conclude that

$$(3.11) \quad \frac{k_2}{k_1} \mathcal{N}(\tilde{x}_i^U, \tilde{\sigma}_i^U) \geq p(x_t | x_{\tau_{i-1}}, y_{\tau_{i-1}}^{\tau_i}) \geq \frac{k_1}{k_2} \mathcal{N}(\tilde{x}_i^L, \tilde{\sigma}_i^L).$$

Therefore, using (3.11), we have

$$\begin{aligned} & E_{u_i \in \sigma(y_{\tau_{i-1}}^{\tau_i}) \vee \mathcal{F}_{\tau_{i-1}}^*} \inf P(|x_{\tau_i} - u_{\tau_i}| \geq c | y_{\tau_{i-1}}^{\tau_i}) \\ & \geq \frac{k_1}{k_2} E_{u_i \in \sigma(y_{\tau_{i-1}}^{\tau_i}) \vee \mathcal{F}_{\tau_{i-1}}^*} \int_{|x - u_{\tau_i}| > c} \mathcal{N}_x(\tilde{x}_i^L, \tilde{\sigma}_i^L) dx \\ & \geq \frac{k_1}{k_2} \int_{|x| > c} \mathcal{N}_x(0, \tilde{\sigma}_i^L) dx, \end{aligned}$$

which implies, by a direct computation using (3.11), that

$$\begin{aligned} (3.12) \quad & E_{u_i \in \sigma(y_{\tau_{i-1}}^{\tau_i}) \vee \mathcal{F}_{\tau_{i-1}}^*} \inf P(|x_{\tau_i} - u_{\tau_i}| \geq c | y_{\tau_{i-1}}^{\tau_i}, x_{\tau_{i-1}}) \\ & \geq \mu \exp\left(\frac{-c^2 h}{2\epsilon \tanh(h\Delta t/\epsilon)}\right) \geq \tilde{\mu} e^{-c^2 h/2\epsilon}, \end{aligned}$$

where $\mu, \tilde{\mu}$ are deterministic constants independent of Δt and ϵ . Therefore, substituting in (3.6), we have that

$$(3.13) \quad J_\pi \leq \left(1 - \mu \exp\left(-\frac{c^2 h}{2\epsilon \tanh(h\Delta t/\epsilon)}\right)\right)^{n(\epsilon)}.$$

We may optimize $n(\epsilon)$ to get the best possible bound in (3.13) as a function of ϵ . For our needs, it suffices to take $n(\epsilon) = 1$ and conclude that

$$\lim_{\epsilon \rightarrow 0} \epsilon \log \inf_{u \in \Pi} P^u(\|x - u\|_T > c) \geq -\frac{c^2 h}{2},$$

and (3.1) is proven. To see (3.1a), take $\Delta t = 1$ and note that as above the performance is bounded by the performance of the best controller with the additional information $x_{\tau_{i-1}}$. Due to the Markov structure and the uniform bound on the escape probability over an interval of length 1 (here, we use the fact that the escape probabilities are not only bounded below as in the proof above but are actually also bounded above with the same exponent, again due to (3.11)), we have that

$$E(\tau) \leq E \sum_{t=1}^{\infty} t P(|x_{\tau_t} - u_{\tau_t}| > c | x_{\tau_{t-1}}) \prod_{s=1}^{t-1} P(|x_{\tau_s} - u_{\tau_s}| \leq c | x_{\tau_{s-1}})$$

$$\leq k e^{-c^2 h/2\epsilon} \sum_{t=1}^{\infty} t(1 - k e^{-c^2 h/2\epsilon})^{t-1} \leq \tilde{k} e^{c^2 h/2\epsilon},$$

where k, \tilde{k} are ϵ independent constants, and (3.1a) follows. \square

Proof of Lemma 3.2. Define $\mu_t \triangleq u_t - x_t$, we have that

$$(3.14) \quad d\mu_t = (m(\hat{u}_t) - m(x_t)) dt - \frac{h}{\epsilon} \mu_t dt + \sqrt{2} d\tilde{w}_t,$$

where $\tilde{w}_t \triangleq (v_t - w_t)/\sqrt{2}$ is a standard Brownian motion. Consider now the equation

$$(3.15) \quad d\tilde{\mu}_t = -\frac{h}{\epsilon} \tilde{\mu}_t dt + \sqrt{2} d\tilde{w}_t,$$

with the same initial conditions as (3.14), i.e., $\tilde{\mu}_0 = \mu_0 = 0$. Setting $\rho_t = \mu_t - \tilde{\mu}_t$ and $\alpha_t = m(\hat{u}_t) - m(x_t)$, we have

$$\frac{d\rho_t}{dt} = \alpha_t - \frac{h}{\epsilon} \rho_t, \quad \rho_0 = 0$$

which yields

$$\rho_t = \int_0^t \alpha_s \exp\left(\frac{h}{\epsilon}(s-t)\right) ds.$$

Since $m(\cdot)$ was assumed bounded, $\alpha(\cdot)$ is bounded and therefore

$$(3.16) \quad \|\rho\|_t \leq \frac{M\epsilon}{h}$$

for all $0 \leq t \leq \infty$. Therefore, for any time interval, the large deviations behavior of (3.14) is the same as that of (3.15). Note that $\tilde{\mu}_t$ is a Gaussian process, and $\tilde{\mu}_t = \sqrt{2} \int_0^t \exp((h/\epsilon)(s-t)) d\tilde{w}_s$. Therefore,

$$E\tilde{\mu}_{t_1}\tilde{\mu}_{t_2} = \frac{\epsilon}{h} \left[\exp\left(-\frac{h}{\epsilon}|t_1 - t_2|\right) - \exp\left(-\frac{h}{\epsilon}(t_1 + t_2)\right) \right].$$

By known results for the extrema of Gaussian processes, cf. [15],

$$(3.17) \quad \lim_{\epsilon \rightarrow 0} \frac{P(\|\tilde{\mu}\|_T > c)}{(h/\sqrt{2\pi\epsilon}) \int_c^\infty \exp(-(hx^2/2\epsilon)) dx} = 1,$$

which is more than sufficient to show (3.2). Since $\tilde{\mu}_t$ is a Markov process, by (3.17) we obtain (3.2a) by following the derivation of [16, Chap. 6]. \square

Lemmas 3.1 and 3.2 jointly demonstrate that the filter \hat{u}_t is asymptotically logarithmically optimal. Following are some related issues and generalizations.

First, note that the performance is monotoneous in the value of ϵ . To see that, adjunct to y_t a new additional observation $d\tilde{y}_t = \tilde{\epsilon} d\tilde{v}_t$ where \tilde{v}_t is independent of all other Brownian motions in the model. Clearly, the optimal control does not change, and the optimal value is not worse than when observing the projection

$$\begin{aligned} \hat{y}_t = y_t + \tilde{y}_t &= \int_0^t h(x_t - u_t) dt + \epsilon v_t + \tilde{\epsilon} \tilde{v}_t \\ &= \int_0^t h(x_t - u_t) dt + \sqrt{\epsilon^2 + \tilde{\epsilon}^2} v_t, \end{aligned}$$

where $\stackrel{d}{=}$ denotes equality in distribution, which proves the monotonicity in ϵ . On the other hand, the problem of successfully designing an $h(x - u)$ that is optimal under a specific constraint on h (e.g. $|h|_{\max}$ or $\int_{-c}^c h(\theta)d\theta$) remains open.

Next, we comment on generalizations of the results in this section: A version of the upper bound on the performance (Lemma 3.2) is still valid under a general model ($\sigma(x)$, multidimensional x , nonlinear h) since it can be based on general large deviation estimates instead of the Gaussian estimates we have used. However, the lower bound (3.1) is much harder to compute: for x multidimensional, $h(\cdot)$ linear in the interior of its support, of rank k , $\sigma(x) \equiv \sigma$ and rank $\sigma \sigma^T = k$, one may still apply the results of [18, §4], and obtain the analog of Lemma 3.1. In general, however, Lemma 3.1 depends on obtaining good large deviation bounds on the conditional distribution. For σ or h nonlinear, no such results are available even in the one-dimensional case, and the best one has is [19] or [14], which is not good enough, for it does not provide an exponential rate of decay with ϵ for the unnormalized density. The general multidimensional case (rank $h < k$) is even more difficult since then the conditional density might not converge (as $\epsilon \rightarrow 0$) weakly to a Dirac measure as it does in the one-dimensional case.

Appendix. The discrete time case. In this appendix, we consider the discrete time stochastic control problem associated with the tracking problem.

Let Σ be the state space $\Sigma = \mathbb{R}^k$ or $\Sigma = \{1, 2, \dots, k\}$. Throughout, the norm on Σ is denoted $|\cdot|$. Let x_n be a discrete time Markov chain taking values in Σ with transition probability $P_{x_{n+1}}(\cdot | x_n)$. Let $h: \Sigma \rightarrow \mathbb{R}^1$ be a measurable function of compact support, and define the observation $y_n \in \mathbb{R}^1$ as

$$(A.1) \quad y_{n+1} = y_n + h(x_{n+1} - u_{n+1}) + \epsilon v_{n+1}, \quad y_0 = 0,$$

where $\epsilon > 0$, $\{v_n\}$ is a sequence of independent and identically distributed standard Normal random variables, and $u_{n+1} \in \Sigma$ is an admissible control, i.e., a measurable map $u_{n+1}: \{y_0, y_1, \dots, y_n\} \rightarrow \Sigma$. To simplify matters, the following restrictive assumption is made throughout this appendix:

(A) u_n takes values in a compact convex set $U \subset \Sigma$.

Dispensing of (A) is standard via tightness arguments and will therefore not be considered here. The class of admissible controls will be denoted Π in the sequel. The optimal control problem considered here is to find a policy $\{u_n\}$ that minimizes the cost

$$(A.2) \quad J^u \triangleq \text{Prob}(\|x - u\|_N > c),$$

where c is a given constant > 0 and $\|a\|_n = \sup_{0 \leq t \leq n} |a_t|$.

Remark. In general, we should allow for randomized controls u . However, using the results of [3], we can show that the infimum of J^u over randomized controls is the same as the infimum of J^u over strict sense controls. Since we will show below that the infimum over strict sense controls is indeed achieved, it is enough to consider strict sense controls only. That is in contrast with the continuous time case.

We now transfer this problem to the standard partially observed framework. First, enlarge the state space Σ to

$$\tilde{\Sigma} = \Sigma \times \{0, 1\}.$$

Next, define a *controlled Markov* chain on $\tilde{\Sigma}$ by considering pairs $\mu_n^u = \{x_n, z_n\}$, where x_n is as before and $z_1 = 0$, $z_n = 1_{\|x - u\|_n \geq c}$. Note that z_n has at most one transition

from zero to one. Furthermore, μ_n^u has a transition kernel of the form

$$\begin{aligned}
 (A.3) \quad & P_{\mu_{n+1}^u}((A, 0) \mid \mu_n^u = (x_n, 0)) = \int_A dP_{x_{n+1}}(\theta \mid x_n) 1_{|\theta - u_{n+1}| \leq c}, \\
 & P_{\mu_{n+1}^u}((A, 1) \mid \mu_n^u = (x_n, 0)) = \int_A dP_{x_{n+1}}(\theta \mid x_n) 1_{|\theta - u_{n+1}| > c}, \\
 & P_{\mu_{n+1}^u}((A, 0) \mid \mu_n^u = (x_n, 1)) = 0, \\
 & P_{\mu_{n+1}^u}((A, 1) \mid \mu_n^u = (x_n, 1)) = \int_A dP_{x_{n+1}}(\theta \mid x_n).
 \end{aligned}$$

Let $M_1(\tilde{\Sigma})$ denote the space of probability measures on $\tilde{\Sigma}$. $M_1(\tilde{\Sigma})$ is a Polish space, with the Levy-Prohorov metric, which is compatible with weak convergence [8, Chap. 3, Thm. 1.7]. We take $M_1(\tilde{\Sigma})$ to be our new state space, and on $M_1(\tilde{\Sigma})$ we will consider the process of conditional measures of μ_n^u conditioned on $y_o^n = \{y_o, y_1, \dots, y_n\}$,

$$Q_{\mu_n}^u(\cdot) \triangleq P_{\mu_n^u}(\cdot \mid y_o^n).$$

Note that for all Borel sets $A \subset B(\tilde{\Sigma})$,

$$\begin{aligned}
 (A.4) \quad & Q_{\mu_{n+1}}^u(A) = \sum_{\theta=0}^1 \int_{\eta \in \Sigma} dQ_{\mu_n}^u(\eta, \theta) P_{\mu_{n+1}^u}(A \mid \mu_n = \{\eta, \theta\}) \\
 & \frac{1}{\sqrt{2\pi\epsilon}} e^{-(y_{n+1} - y_n - h(\eta - u_{n+1}))^2 / 2\epsilon^2}.
 \end{aligned}$$

$Q_{\mu_{n+1}}^u \in M_1(\tilde{\Sigma})$ will be our new state. By the innovation decomposition,

$$(A.5) \quad y_{n+1} - y_n = \sum_{\theta=0}^1 \int_{\eta \in \Sigma} dQ_{\mu_n}^u(\eta, \theta) h(\eta - u_{n+1}) + \epsilon \hat{v}_{n+1},$$

where $\{\hat{v}\}_n$ is an independent and identically distributed standard Normal sequence such that \hat{v}_n is independent of $(Q_{\mu_1}^u, \dots, Q_{\mu_n}^u)$. Consequently, combining (A.4) and (A.5), one sees that the conditional measures form a controlled Markov chain in the sense of [3]. Moreover, note that

$$J^u = E(1_{\|x-u\|_N > c}) = \int_{\Sigma} dP_{\mu_N}(\theta, 1) = E \int_{\Sigma} dQ_{\mu_N}^u(\theta, 1),$$

where the last expectation is with respect to the sequence $\{\hat{v}_i\}$, $i = 1, \dots, N$. Therefore, the original problem transforms into a control problem in the variables $Q_{\mu_n}^u$ with terminal cost.

As a last general remark, note that

$$J^u = 1 - E \int_{\Sigma} dQ_{\mu_N}^u(\theta, 0).$$

Moreover, as far as the optimization problem is concerned, modifying $h(x_n - u_n)$ to $\tilde{h}(x_n - u_n, z_n)$ with $\tilde{h}(x_n - u_n, 0) = h(x_n - u_n)$ and $\tilde{h}(x_n - u_n, 1) = 0$ will not change the value function for any y measurable control. Indeed, let

$$\tilde{y}_{n+1} = \tilde{y}_n + \tilde{h}(x_{n+1} - u_{n+1}, z_{n+1}) + \epsilon v_n.$$

Let $u_n(\cdot)$ be an admissible control. Up to $\tau = \inf\{t \mid z_t = 1\}$, $y_n = \tilde{y}_n$, and therefore, $J_N^u = J_N$. Further,

$$(A.6) \quad \tilde{y}_{n+1} - \tilde{y}_n = \int_{\eta \in \Sigma} dQ_{\mu_n}^u(\eta, 0) h(\eta - u_{n+1}) + \epsilon \hat{v}_n.$$

As a result of the above, and using (A.4)–(A.6), we may rewrite the optimization problem in terms of $\tilde{Q}_{x_n}(A) \triangleq Q_{x_n}(A, 0)$ as follows.

(P) PROBLEM DEFINITION.

Maximize $E \int_{\Sigma} d\tilde{Q}_{x_n}(\theta)$ subject to the state equations (in $M_1(\Sigma)$):

$$(A.7) \quad \tilde{Q}_{x_{n+1}}(A) = \int_{\eta \in \Sigma \cap \{|x - u_{n+1}| \leq c\}} d\tilde{Q}_{x_n}(\eta) P_{x_{n+1}}(A | x_n = \eta) \\ \cdot \frac{1}{\sqrt{2\pi\epsilon}} \exp \left(- \left(\int_{\Sigma \cap \{|\theta - u_{n+1}| \leq c\}} d\tilde{Q}_{x_n}(\theta) h(\theta - u_{n+1}) - h(\eta - u_{n+1}) + \epsilon \hat{v}_n \right)^2 / 2\epsilon^2 \right),$$

where u_{n+1} is a measurable map of $\{\tilde{Q}_{x_n}\}$.

To discuss existence and structural results, we distinguish between the following two cases:

(a) $\Sigma = \mathbb{R}^k$, U is a closed compact set in \mathbb{R}^k .

(b) $U = \Sigma = \{1, 2, \dots, k\}$, and we take the discrete topology on both.

(a) THE CONTINUOUS CASE. We make the following assumption on the transition probability $P_{x_{n+1}}(A | x_n)$:

(A-C) $P_{x_{n+1}}(A | x_n)$ possesses a continuous density with respect to Lebesgue measure, which we denote by $P_{x_{n+1}}(\theta | x_n)$.

Using an induction argument, we have immediately the following Lemma.

LEMMA A.1. Assume (A-C). Then $\tilde{Q}_{x_{n+1}}^u(A)$ possesses a density with respect to Lebesgue measure and is a continuous function of u_1, u_2, \dots, u_n .

Equipped with Lemma A.1, we may apply a version of [3, Prop. 8.6] to show Theorem A.1.

THEOREM A.1. An optimal Markovian, nonrandomized control for the problem (P) exists, and is given by the solution to the dynamic programming equation:

$$V(m, \tilde{Q}) = \sup_u E^u(V(m+1, \tilde{Q}^*))$$

with the terminal condition

$$V(N, \tilde{Q}) = \int_{\Sigma} d\tilde{Q}(\theta)$$

where E^u is with respect to the transition kernel given by (A.7), i.e., expectation with respect to the Gaussian random variable \hat{v} in the formula

$$\tilde{Q}^*(A) = \frac{1}{\sqrt{2\pi\epsilon}} \int_{\eta \in \Sigma \cap \{|x-u| \leq c\}} d\tilde{Q}(\eta) P(A|\eta) \exp \left(- \left(\int_{\Sigma \cap \{|\theta-u| \leq c\}} d\tilde{Q}(\theta) h(\theta-u) - h(\eta-u) + \epsilon \hat{v} \right)^2 / 2\epsilon^2 \right).$$

Remark. The results of this section can be extended to the case where U is not compact, by noting that $P(\|x\| > K) \xrightarrow{K \rightarrow \infty} 0$, $P(\|\theta\| > K) \xrightarrow{K \rightarrow \infty} 0$ and by using the tightness of P_x , P_y , which follows from those relations.

(b) THE DISCRETE CASE. In this case, $\Sigma \in \{1, \dots, k\}$ and we take $u = \{1, \dots, k\}$. Note that the conditional density $P^u(x_i|y_o^i)$ is now a vector in \mathbb{R}^k . The problem (P) is now identical to the one solved in [7, §8.2]. The results of [7, §8.2] imply that the dynamic programming equation for the problem (P) yields an optimal measurable control, with the dynamic programming equation taking the form

$$(A.8a) \quad V(n, \tilde{Q}) = \sup_u E^u(V(n+1, \tilde{Q}^*))$$

$$(A.8b) \quad V(N, \tilde{Q}) = \sum_{i=1}^k \tilde{Q}(i)$$

where

$$\tilde{Q}^*(j) = \sum_{\ell \in \{1, \dots, k\} \cap \{|\ell-u| \leq c\}} \tilde{Q}(\ell)$$

We conclude this section by noting that solving (P) is equivalent to the following iterative five-step procedure:

- (a) Compute $P_{x_n}(A|y_o^{n+1}, \|x-u\|_n < c)$.
- (b) Propagate $P_{x_n}(A|y_o^{n+1}, \|x-u\|_{n+1} < c)$ according to the dynamics of x_n , i.e., form

$$P_{x_{n+1}}(A|y_o^{n+1}, \|x-u\|_n < c) = \int P_{x_n}(d\eta|y_o^{n+1}, \|x-u\|_n < c) P_{x_{n+1}}(A|\eta).$$

- (c) Choose a u_{n+1} according to an (optimal or suboptimal) strategy.

- (d) Compute $P_{x_{n+1}}(A|y_o^{n+1}, \|x-u\|_{n+1} < c)$.

- (e) Observe y_{n+2} and compute $P_{x_{n+1}}(A|y_o^{n+2}, \|x-u\|_{n+1} < c)$.

Intuitive candidates for suboptimal rules in step c above would be to pick up u_{n+1} as one of the followings:

- (R1) The conditional mean of $P_{x_{n+1}}(\cdot|y_o^{n+1}, \|x-u\|_n < c)$.
- (R2) The maximizer of the density $P_{x_{n+1}}(\cdot|y_o^{n+1}, \|x-u\|_n < c)$ (if such density and maximizer exist).
- (R3) The one-step best predictor, i.e.,

$$\tilde{u}_{n+1} = \arg \max_u \int_{|\eta-u| \leq c} dP_{x_{n+1}}(\eta|y_o^{n+1}, \|x-u\|_n < c)$$

The fact that the different rules may generally lead to very different controls is obvious. When $\epsilon \rightarrow 0$, i.e., the observation noise is small, as seen in §3, (R1) is asymptotically optimal (for the continuous model, although the results are similar in the discrete case), moreover (R2) and (R3) yield the same asymptotic performance as (R1).

REFERENCES

- [1] Y. BAR-SHALOM AND T. E. FORTMAN, *Tracking and Data Association*, Academic Press, 1988.
- [2] V. BORKAR, *Existence of optimal controls for partially observed diffusions*, Stochastics, 11 (1983), pp. 103–142.
- [3] D. P. BERTSEKAS AND S. E. SHREVE, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.
- [4] M. BISMUT, *Partially observed diffusions and their control*, SIAM J. Control Optim., 20 (1982), pp. 302–309.
- [5] R. COHEN AND G. MAZIOTTO, *Stochastic control of partially observed systems via impulse control problems*, Stochastics, 26 (1989), pp. 101–127.
- [6] P. DUPUIS AND H. J. KUSHNER, *Minimizing escape probabilities: a large deviations approach*, SIAM J. Control Optim., 27 (1989), pp. 432–445.
- [7] E. B. DYNKIN AND A. A. YUSHKEVITSH, *Controlled Markov processes*, Springer-Verlag, Berlin, New York, 1979.
- [8] S. N. ETHIER AND T. G. KURTZ, *Markov Processes: Characterization and Convergence*, Wiley, New York, 1986.
- [9] W. FLEMING AND E. PARDOUX, *Optimal control for partially observed diffusions*, SIAM J. Control Optim., 20 (1982), pp. 261–285.
- [10] R. S. LIPTSER AND A. N. SHIRYAYEV, *Statistics of Random Processes I*, Springer-Verlag, Berlin, New York, 1977.
- [11] S. M. MEERKOV AND T. RUNOLFSON, *Residence time control*, IEEE Trans. Automat. Control, 33 (1988), pp. 323–332.
- [12] ———, *Output aiming control*, Proc. 26th IEEE Conf. on Dec. Cont., Los Angeles CA, 1987, pp. 1734–1739.
- [13] E. PARDOUX, *Filtrage de diffusion avec conditions frontieres: caracterisation de la densite conditionnelle*, in Journees de statistique des processus aleatoire, Lecture Notes in Mathematics D. Dachuna-Castelle and B. Van Cutsem, eds., Springer-Verlag, Berlin, New York, pp. 163–188.
- [14] J. PICARD, *Nonlinear filtering of one-dimensional diffusions in the case of high signal to noise ratio*, SIAM J. Appl. Math., 46 (1986), pp. 1098–1125.
- [15] M. TALAGRAND, *Small tails for the supremum of a Gaussian process*, Ann. Inst. Henri Poincare, 24 (1988), pp. 307–315.
- [16] S. R. S. VARADHAN, *Large Deviations and Applications*, Society for Industrial and Applied Mathematics, Philadelphia, 1984.
- [17] M. ZAKAI, *On the optimal filtering of diffusion processes*, Z. Wahr. Ver. Geb., 11 (1969), pp. 230–243.
- [18] O. ZEITOUNI, *Approximate and limit results for nonlinear filters with small observation noise: the linear sensor and constant diffusion coefficient case*, IEEE Trans. Automat. Control, 33 (1988), pp. 595–599.
- [19] ———, *Limit results for nonlinear filters with small observation noise: the general one dimensional case*, in Analysis and Control of Nonlinear System (Proc. MTNS 1987), C. I. Byrnes, C. F. Martin, and R. E. Saeks, eds., North Holland, New York, 1988, pp. 271–277.
- [20] O. ZEITOUNI AND M. ZAKAI, *On the optimal tracking problem*, EE Pub. Number 767, Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa, Israel, 1990.

A MONTE CARLO METHOD FOR SENSITIVITY ANALYSIS AND PARAMETRIC OPTIMIZATION OF NONLINEAR STOCHASTIC SYSTEMS: THE ERGODIC CASE.*

HAROLD J. KUSHNER^{†‡} AND JICHUAN YANG^{†§}

Abstract. For high-dimensional or nonlinear problems there are serious limitations on the power of available computational methods for the optimization or parametric optimization of stochastic systems of diffusion type. The paper develops an effective Monte Carlo method for obtaining good estimators of systems sensitivities with respect to system parameters, when the system is of interest over a long period of time. The value of the method is borne out by numerical experiments, and the computational requirements are favorable with respect to competing methods when the dimension is high or the nonlinearities “severe.” The method is a type of “derivative of likelihood ratio” method. For a wide class of problems, the cost function or dynamics need not be smooth in the state variables; for example, where the cost is the probability of an event or “sign” functions appear in the dynamics. Under appropriate conditions, it is shown that the invariant measures are differentiable with respect to the parameters. Since the basic diffusion (or other) model cannot be simulated exactly, simulatable approximations are discussed in detail, and estimators of the derivatives of the cost functions for these approximations are obtained and analyzed. It is shown that these estimators and their expectations converge to those for the original problem. Thus, we prove a robustness result for the sensitivity estimators, namely that the derivatives of the ergodic cost functions (and their estimators) for the simulatable approximations converge to those for the approximated process. Such results are essential if a simulation based method is to be used with confidence.

Key words. Monte Carlo method for diffusions, parametric optimization of stochastic systems, sensitivity analysis, optimization of stochastic systems, nonlinear stochastic systems, high-dimensional stochastic systems, parametric optimization of diffusion processes, likelihood ratio method for sensitivity analysis, parametric derivatives of invariant measures, ergodic control

AMS(MOS) subject classifications. 62E25, 93E20, 93E25

1. Introduction. This paper is concerned with a key question in the use of recursive Monte Carlo methods for system optimization, when the system operation and cost are of interest for a long period of time. For many control systems, the control is given a priori in a parametrized form, and for the use of Monte Carlo methods for the optimization of the parameter, we need good estimators of the derivatives of the cost function with respect to the parameter.

Reference [1] develops a very useful method for doing this when the system is of the diffusion or related type and the control interval of concern is finite. Numerical approximations to the unbiased estimators were developed and analyzed, and simulations showed that the method can be superior to competing methods if the system dimension is large or if the system is nonlinear. In this paper, the results of [1] are extended to the ergodic cost problem. New difficulties arise, since we essentially need to deal with derivatives of the invariant measures with respect to the control parameters and with the convergence of suitable computable approximations. Owing to these “ergodic” problems, the assumptions are stronger here than in [1].

* Received by the editors September 27, 1990; accepted for publication (in revised form) March 7, 1991.

[†] Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912.

[‡] The work of this author was partially supported by Air Force Office of Scientific Research grant 89-0015 and Army Research Office grant DAAL-03-86K-0171.

[§] The work of this author was partially supported by National Science Foundation grant ECS-8913351.

Let $x(\cdot)$ be defined by the diffusion

$$(1.1) \quad dx = b(x, \alpha)dt + \sigma(x)dw, \quad x \in R^r,$$

where $a(x) = \sigma(x)\sigma'(x)$ is nondegenerate and α is a control parameter to be chosen. For each α of interest, let $x(\cdot)$ have a unique invariant measure $\mu(\alpha)$. Precise conditions will be given below. For “smooth cost rate” $k(\cdot, \alpha)$, define the “ergodic cost”

$$(1.2) \quad \langle \mu(\alpha), k(\alpha) \rangle \equiv \int \mu(dx, \alpha)k(x, \alpha) \equiv \bar{k}(\alpha).$$

We wish to get an unbiased estimator of $\partial \bar{k}(\alpha)/\partial \alpha$ (as well as reasonable “numerical” approximations from sample simulations) at selected values of α . Such estimators are necessary if we wish to minimize $\bar{k}(\alpha)$ over α by some recursive Monte Carlo (stochastic approximation) method.

Control problems are frequently of this type; i.e., the control is given in a parametric form. Often, a full optimal feedback control is not desired since it might be very difficult to implement, and all the state variables are not available. A good class of parametrized controls might be known, however. See [1] for some examples and further motivation, as well as a discussion of alternative approaches.

Generally, we cannot easily evaluate $\bar{k}(\alpha)$ or its derivatives. Then we might seek a method for getting good estimators that can be used in a recursive Monte Carlo optimization method. The ease of getting the estimates and their quality are key issues in such an approach. The estimators are to be obtained by simulations of (1.1) or of approximations to (1.1), since the solution of (1.1) cannot be known exactly.

Reference [1] developed a general “likelihood ratio derivative”-based method for getting such estimators, under conditions that are much broader than those used in this paper, but for a “finite time” problem. The numerical data in [1], and that obtained subsequently, show that the method can be quite superior to its competitors for nonlinear and high-dimensional systems. The quality of the estimator is judged by the “variance per CPU time required in the simulations.” The reader is referred to [1] for more motivation and examples. The ergodic cost problem is more difficult and requires stronger (hence, the nondegeneracy) conditions. Actually, the method has been successfully tested on many degenerate problems of the type used in [1], so that the conditions that our analysis requires can undoubtedly be weakened. There are ready extensions to the jump-diffusion, reflection, and other standard models. To introduce the idea, we give a brief informal review of one idea in [1], but we use slightly different terminology and stronger conditions than those used in [1].

For given $T < \infty$, define the “finite time” costs

$$C(x, \alpha) = \int_0^T k(x(s), \alpha)ds + k_0(x(T), \alpha), \quad \bar{C}(x, \alpha) = E_x^\alpha C(x, \alpha),$$

where E_x^α denotes the expectation with parameter α and $x(0) = x$. We always use α_0 to denote the point at which the derivative is to be taken. With no loss of generality, α will be a real number, since for the vector case we can estimate the derivative for each component separately. Let $P_x^\alpha(T)$ denote the measure induced by the solution to (1.1) with the initial condition $x(0) = x$, on $C^r[0, T]$, the space of R^r -valued continuous functions on $[0, T]$, with the sup norm. Let $b(x, \alpha)$, $k(x, \alpha)$, and $k_0(x, \alpha)$ be α -differentiable, and define $\alpha = \alpha_0 + \delta\alpha$ and $\delta b(x, \alpha_0, \delta\alpha) = b(x, \alpha_0 + \delta\alpha) - b(x, \alpha_0)$.

Define

$$\begin{aligned}\xi(0, T; \alpha_0, \delta\alpha) &= \int_0^T [\sigma^{-1}(x(s))\delta b(x(s), \alpha_0, \delta\alpha)]' dw(s) \\ &\quad - \frac{1}{2} \int_0^T |\sigma^{-1}(x(s))\delta b(x(s), \alpha_0, \delta\alpha)|^2 ds,\end{aligned}$$

and the Radon–Nikodym derivative

$$(1.3) \quad \frac{dP_x^{\alpha_0 + \delta\alpha}(T)}{dP_x^{\alpha_0}(T)} = \exp \xi(0, T; \alpha_0, \delta\alpha).$$

Define $Z(\cdot, \alpha_0)$ by

$$\begin{aligned}(1.4) \quad Z(T, \alpha_0) &= \int_0^T [\sigma^{-1}(x(s))b_\alpha(x(s), \alpha_0)]' dw(s) \\ &= \int_0^T [b'_\alpha(x(s), \alpha_0)a^{-1}(x(s))][dx(s) - b(x(s), \alpha_0)ds].\end{aligned}$$

We use the subscripted $b_\alpha(x, \alpha_0)$ and others to denote the α -derivatives at α_0 . Then the quantities

$$\begin{aligned}(1.5) \quad Q(\alpha_0) &= \int_0^T [k(x(s), \alpha_0)Z(s, \alpha_0) + k_\alpha(x(s), \alpha_0)]ds \\ &\quad + k_0(x(T), \alpha_0)Z(T, \alpha_0) + k_{0,\alpha}(x(T), \alpha_0),\end{aligned}$$

$$\begin{aligned}(1.5') \quad \hat{Q}(\alpha_0) &= \int_0^T [(k(x(s), \alpha_0) - \bar{k}(x(s), \alpha_0))Z(s, \alpha_0) + k_\alpha(x(s), \alpha_0)]ds \\ &\quad + (k_0(x(T), \alpha_0) - \bar{k}_0(x(T), \alpha_0))Z(T, \alpha_0) + k_{0,\alpha}(x(T), \alpha_0),\end{aligned}$$

where we use $\bar{k}(x(s), \alpha_0) = E_x^{\alpha_0} k(x(s), \alpha_0)$, are unbiased estimators of $\bar{C}_\alpha(x, \alpha_0)$. Thus, if a path of $x(\cdot)$ is available, one can calculate or approximate (1.5) or (1.5').

To avoid the very time-consuming task of evaluating (from the simulations) $\bar{k}(x(s), \alpha_0)$ for each $s \leq T$, in (1.5'), we usually use $\bar{k}(x(T), \alpha_0)$ in place of $\bar{k}(x(s), \alpha_0)$, and with good results.

Generally, paths of the true model $x(\cdot)$ are not available, and we can only approximate via a numerical method (say, a discrete time approximation). Reference [1] discusses two basic classes of such approximations and proves that the estimators obtained from them are good. Getting good estimators is more difficult for the ergodic problem, since we also need to truncate the infinite time interval and approximate (at least implicitly) derivatives of invariant measures, a nontrivial problem.

The proofs use a representation of the invariant measure of the diffusion process in terms of that of an imbedded Markov chain, defined by the random return times to a “recurrence set,” as well as certain Girsanov transformations defined on these “return intervals.” To be sure that these transformations are well defined, a bound on an exponential moment of the return time is needed. This is provided by the stability result in §2. Section 3 is concerned with ergodic properties of the diffusion

model. The imbedded Markov chain is defined, the invariant measure of the diffusion is defined in terms of this Markov chain, and the needed recurrence (ϕ -recurrence) properties of the chains are stated. Section 4 is concerned with the existence of the derivative of the invariant measure of the diffusion with respect to the parameter. The differentiability is first shown for the invariant measure of the imbedded chain, and this is then used to obtain the result for the diffusion. The differentiability is in two senses: setwise convergence and weak convergence. Some preliminary results concerning equicontinuity of certain sets of functions and invertibility of the operator $I - \tilde{P}(\alpha_0)$ (defined in the section) are first proved. It is also shown that the derivative of the invariant measure can be well approximated by the derivative of the transition function for large enough time.

Since the diffusion model is an "ideal" model and the paths can at best be approximated in some statistical sense, we need to know that the natural approximations can be used with confidence in any implementation. Reference [1] deals with two types of approximations, a discrete time model and a Markov chain approximation. Either can be used here, but we restrict our attention to the first approximation. The model is introduced in §5, and some preliminary sensitivity results are stated there. Some needed stability estimates (analogous to the estimates of §2), uniform in the approximation parameter, are obtained in §6. The main theoretical results for the approximations are in §7, where, after getting some preliminary results concerning the rate of convergence of certain quantities to their "invariant means," it is shown that the invariant measure of the discrete time approximation is differentiable with respect to the control parameter and that the derivatives converge to the derivative of the invariant measure of the diffusion. Results concerning finite time approximations are also shown. The results imply an important robustness of the derivatives with respect to the model. This is new and very useful from the point of view of applications, since otherwise general results concerning the existence of the derivatives for the ideal model would not have much practical relevance.

Numerical data is given in §8. The basic method of implementation requires the use of a discrete parameter approximation, over a finite time period. The period needs to be large enough to capture the "ergodic effects." Two methods are compared; a finite difference method, which has been altered to be fairly efficient, and several forms of our method. The comparison depends on the problem, but it is clear that for a large class of nonlinear problems, our method is preferable. We note that reasonable examples can be constructed so that any chosen method works best, keeping an open mind in any application.

The analysis is restricted to nondegenerate diffusion models, but a similar analysis can be carried out with various related process, provided only that ergodic results analogous to those of §3 are available.

2. Stability of $x(\cdot)$. To develop the ergodic results and use a Girsanov measure transformation method on random unbounded intervals, suitable stability properties of $x(\cdot)$ must be proved. We will use the following assumptions. The parameter α will be confined to a compact interval A_0 with α_0 in its interior. The symbol $'$ denotes transpose.

Assumption 2.1. $b(\cdot, \cdot)$ and $\sigma(\cdot)$ are continuous, $\sigma(\cdot)$ is bounded, and $\sigma(x)\sigma'(x) = a(x) \geq \varepsilon_0 I$ for some $\varepsilon_0 > 0$. For some $K < \infty$, $|b(x, \alpha)| \leq K|x| + K$.

Assumption 2.2. Equation (1.1) has a unique weak sense solution for each $x(0) = x$ and $\alpha \in A_0$.

Assumption 2.3. There is a twice continuously differentiable Lyapunov function $0 \leq V(x) \rightarrow \infty$ as $|x| \rightarrow \infty$ and $\varepsilon_1 > 0$ such that

- (a) $V_{xx}(x)$ is bounded and continuous,
- (b) $V'_x(x)b(x, \alpha) \leq -\varepsilon_1 < 0$ for large $|x|$, $\alpha \in A_0$,
- (c) $\lim_{|x| \rightarrow \infty} \sup_{\alpha \in A_0} |V_x(x)|^2 / |V'_x(x)b(x, \alpha)| < \infty$,
- (d) $\lim_{|x| \rightarrow \infty} \sup_{\alpha \in A_0} |V_{xx}(x) \cdot a(x)| / |V'_x(x)b(x, \alpha)| < 2$.

Assumption 2.4. When $b(x, \alpha) \equiv 0$, (1.1) has a unique weak sense solution for each $x = x(0)$.

Assumption 2.5. There is a bounded continuous function $b_\alpha(\cdot, \alpha_0)$ such that as $\delta\alpha \rightarrow 0$,

$$\delta b(x, \alpha_0, \delta\alpha) / \delta\alpha \rightarrow b_\alpha(x, \alpha_0)$$

boundedly and uniformly on each compact x -set.

Remark on Assumption 2.3. The condition does not seem to be very restrictive. It holds, in particular, for the linear case $b(x, \alpha) = A(\alpha)x$, where $A(\alpha)$ is “uniformly stable” for $\alpha \in A_0$.

The conditions in Assumption 2.3 can be satisfied when we start with a stable system that satisfies these conditions and then modifies the system by a Girsanov measure transformation. There are existence theorems for stochastic Lyapunov functions. The proofs in [7] can be extended to show the existence of the required $V(\cdot)$ under appropriate conditions.

Remark on Assumption 2.2. Assumption 2.4 and the stability Theorem 2.1 imply Assumption 2.2, but it is useful to isolate it as a separate condition.

THEOREM 2.1. *Suppose that Assumptions 2.1–2.3 hold. There is a compact set Q , which is the closure of its interior such that for each compact $Q_1 \supset Q$ and τ_1 defined by $\tau_1 = \min\{t: x(t) \in Q\}$, we have for small $\rho > 0$*

$$(2.1) \quad \sup_{\alpha \in A_0} \sup_{x \in Q_1 - Q} E_x^\alpha \exp \rho \tau_1 < \infty.$$

Proof. Let \mathcal{L} denote the differential generator of $x(\cdot)$: $\mathcal{L}f(x) = f'_x(x)b(x, \alpha) + \frac{1}{2}\text{trace } f_{xx}(x) \cdot a(x)$. Then

$$\begin{aligned} \mathcal{L}e^{\rho V(x)} &= \rho e^{\rho V(x)} [V'_x(x)b(x, \alpha) \\ &\quad + \rho \text{trace}(V_x(x)V'_x(x)) \cdot a(x)/2 + \text{trace } V_{xx}(x) \cdot a(x)/2]. \end{aligned}$$

Let Q be large enough and ρ small enough such that for $x \notin Q$ (use Assumption 2.3) and some $\lambda > 0$,

$$(2.2) \quad \mathcal{L}e^{\rho V(x)} \leq -\rho\lambda e^{\rho V(x)}.$$

It then follows that for small ρ and $x \notin Q$

$$(2.3) \quad \mathcal{L}[e^{\lambda\rho t} e^{\rho V(x)}] \leq 0.$$

From (2.3), Itô's lemma, and a stopping time argument, it follows that

$$(2.4) \quad E_x^\alpha e^{\lambda\rho\tau_1} \leq E_x^\alpha e^{\lambda\rho\tau_1} e^{\rho V(x(\tau_1))} \leq e^{\rho V(x)}$$

for small ρ and $x = x(0) \notin Q$, which yields the result. \square

COROLLARY. Suppose that Assumptions 2.1–2.3 hold. Let Q and Q_1 be as in the theorem. Define τ to be the first return time of $x(\cdot)$ to Q after hitting ∂Q_1 . Then, for small $\rho > 0$

$$(2.5) \quad \sup_{\alpha \in A_0} \sup_{x \in \partial Q} E_x^\alpha e^{\rho\tau} < \infty.$$

The proof follows from Theorem 2.1 and the nondegeneracy, and is omitted.

3. Ergodic properties of (1.1). By Assumptions 2.1–2.3 and Theorem 2.1, for each $\alpha \in A_0$, $x(\cdot)$ is a recurrent strong Feller process. Let $P(x, t, A \mid \alpha)$ denote the transition function. By [2] and [3], there is a unique invariant measure $\mu(\alpha)$ with $\mu(R^r, \alpha) = 1$ and $P(x, t, A \mid \alpha) \xrightarrow{t} \mu(A, \alpha)$ as $t \rightarrow \infty$, for each Borel A . For $t > 0$, $P(x, t, \cdot \mid \alpha)$ has a bounded and nowhere zero density with respect to Lebesgue measure, and so does $\mu(\alpha)$.

We next state a representation of $\mu(\alpha)$ first used by Khazminskii [2], which is very useful for analysis, largely because it is difficult to work with ergodic problems and to deal with questions concerning convergence to invariant measures when the state space is unbounded.

Let $G_1 \supset G$ be compact sets, each of which is connected and is the closure of its interior. Denote the boundaries by Γ_1 and Γ , respectively, and let G be strictly interior to G_1 . Let Γ and Γ_1 be differentiable. For any $x(0)$, define the following stopping times:

$$\begin{aligned} \tau' &= \inf\{t: x(t) \in \Gamma_1\}, \\ \tau_1 &= \inf\{t: x(t) \in \Gamma\}, \\ \tau'_1 &= \inf\{t > \tau_1: x(t) \in \Gamma_1\}. \end{aligned}$$

For $n > 1$,

$$\tau_n = \inf\{t > \tau'_{n-1}: x(t) \in \Gamma\}, \quad \tau'_n = \inf\{t > \tau_n: x(t) \in \Gamma_1\}.$$

For $x = x(0) \in \Gamma$, we use τ to denote $\tau_2 - \tau_1 = \tau_2$, the canonical “return” time to Γ , after hitting Γ_1 .

By Theorem 2.1, for small $\rho > 0$,

$$(3.1) \quad \sup_{x \in \Gamma, \alpha \in A_0} E_x^\alpha \tau < \infty, \quad \sup_{x \in \Gamma, \alpha \in A_0} E_x^\alpha e^{\rho\tau} < \infty.$$

Let $\alpha \in A_0$. Define the process $\tilde{X}_n = x(\tau_n)$. By [2] and Assumptions 2.1–2.3, $\{\tilde{X}_n\}$ is a recurrent homogeneous Markov chain on Γ . Let $\tilde{P}(x, n, \cdot \mid \alpha)$ denote its transition probability. It has a unique invariant measure $\tilde{\mu}(\alpha)$.

The chain is also defined for initial condition $x = \tilde{X}_0 \in G$. Even though $\tilde{X}_n \in \Gamma$, for $n \geq 1$, it will be useful to use G as the state space in §6 and afterwards, to unify the notation with that for the approximations. The results up to §5 hold with this change.

Define $\tau(A) = \int_0^\tau I_A(x(s))ds$ for Borel sets A . Then we can write ([2], [3])

$$(3.2) \quad \mu(A, \alpha) = \bar{\mu}(A, \alpha) / \bar{\mu}(R^r, \alpha),$$

where

$$\bar{\mu}(A, \alpha) = \int_\Gamma \tilde{\mu}(dx, \alpha) E_x^\alpha \tau(A).$$

Hence, for bounded measurable $f(\cdot)$, we have the representation

$$(3.3) \quad \langle \mu(\alpha), f \rangle = \frac{\int_{\Gamma} \tilde{\mu}(dx, \alpha) E_x^{\alpha} \int_0^{\tau} f(x(s)) ds}{\int_{\Gamma} \tilde{\mu}(dx, \alpha) E_x^{\alpha} \tau}.$$

Equation (3.3) and various approximations to it will be widely used in the following.

Properties of $\{\tilde{X}_n\}$. The chain $\{\tilde{X}_n\}$ on state space Γ is said to be uniformly ϕ -recurrent (for a given measure ϕ on the Borel sets of Γ) if for each Borel $B \subset \Gamma$ with $\phi(B) > 0$

$$P_x^{\alpha} \{\tilde{X}_i \in B, \text{ some } i \leq m\} \rightarrow 1 \text{ as } m \rightarrow \infty;$$

uniformly in $x \in \Gamma$. A sufficient condition [4, p. 29] is that if $\phi(B) > 0$, there exists $n < \infty$, $\varepsilon > 0$ (which can be B -dependent) such that

$$(3.4) \quad P_x^{\alpha} \{\tilde{X}_i \in B, \text{ some } i \leq n\} \geq \varepsilon, \quad \text{for all } x \in \Gamma.$$

If the chain is ϕ -recurrent and a -periodic then there exists $C < \infty$, $\gamma < 1$ such that for Borel sets B

$$(3.5) \quad |P_x^{\alpha} \{\tilde{X}_n \in B\} - \tilde{\mu}(B, \alpha)| \leq C\gamma^n,$$

and for bounded measurable $f(\cdot)$,

$$(3.6) \quad |E_x^{\alpha} f(\tilde{X}_n) - \tilde{f}^{\alpha}| \leq 2C\gamma^n \|f - \tilde{f}^{\alpha}\|,$$

where $\|f\| = \sup_x |f(x)|$ and $\tilde{f}^{\alpha} = \langle \tilde{\mu}(\alpha), f \rangle$.

The next theorem follows from [3, p. 339, proof of Theorem 5.1]. The model in the reference does not explicitly include a parameter α , but it is easily seen from the proof of the cited theorem that the nondegeneracy and the fact that the moment bounds in Theorem 2.1 do not depend on $\alpha \in A_0$ imply that (3.4) is uniform in $\alpha \in A_0$ for some $\varepsilon > 0$. In fact, we can use $n = 1$. Actually, we will only need the result for $\alpha = \alpha_0$.

THEOREM 3.1. *Suppose that Assumptions 2.1–2.3 hold. $\{\tilde{X}_n\}$ is ϕ -recurrent, where ϕ is Lebesgue measure on Γ . The recurrence is uniform in $\alpha \in A_0$ in the sense that the mean recurrence times are bounded uniformly for $\alpha \in A_0$. There are $C < \infty$, $\gamma < 1$ (not depending on $\alpha \in A_0$) such that (3.5) and (3.6) hold.*

It will be seen below (Lemma 4.1) that $\tilde{P}(x, n, B | \alpha)$ is continuous in x , uniformly in α, B . (The continuity is proved in [3], but we give a different proof, since the details to be used will be needed elsewhere in the paper.)

4. The α -derivative of $\tilde{\mu}(\alpha)$ (setwise sense). Let $C(\Gamma)$ denote the set of bounded and continuous functions on Γ , and $C_c(\Gamma)$ the centered functions: $f_1 \in C_c(\Gamma)$ if $f_1 = f - \tilde{f}$ for $f \in C(\Gamma)$, where $\tilde{f} = \langle \tilde{\mu}(\alpha_0), f \rangle$. To prove the differentiability of $\mu(\alpha)$ at α_0 , we first prove that of $\tilde{\mu}(\alpha)$, and then use (3.3).

DEFINITION. $\tilde{\mu}(\alpha)$ is said to be *differentiable at α_0 in the setwise (or weak) sense* if there is a finite signed measure v such that for each Borel set B

$$v(B) = \lim_{\delta\alpha \rightarrow 0} [\tilde{\mu}(B, \alpha_0 + \delta\alpha) - \tilde{\mu}(B, \alpha_0)] / \delta\alpha.$$

$\tilde{\mu}(\alpha)$ is said to be *differentiable at α_0 in the sense of weak convergence (or weak* sense)* if there is a finite signed measure v such that for each $f \in C(\Gamma)$,

$$\langle v, f \rangle = \lim_{\delta\alpha \rightarrow 0} \langle \tilde{\mu}(\alpha_0 + \delta\alpha) - \tilde{\mu}(\alpha_0), f \rangle / \delta\alpha.$$

DEFINITION. Let $L^\infty(\Gamma)$ denote the bounded Borel measurable functions on Γ . For any Borel set H , let $\mathcal{B}(H)$ denote the Borel subsets of H . Define the operator $\tilde{P}(\alpha)$ on $L^\infty(\Gamma)$ by $\tilde{P}(\alpha)f(x) = E_x^\alpha f(\tilde{X}_1)$.

LEMMA 4.1. Suppose that Assumptions 2.1–2.4 hold. Then the set $\{\tilde{P}(\alpha)L^\infty(\Gamma), \alpha \in A\}$ (restricted to functions with $\|f\| \leq 1$) is equicontinuous.

Proof. Define the process $y(\cdot)$ by $y(0) = x$ and

$$(4.1) \quad dy = \sigma(y)dw.$$

Define

$$\xi_0^\alpha(u, v) = \int_u^v [\sigma^{-1}(y(s))b(y(s), \alpha)]' dw(s) - \frac{1}{2} \int_u^v |\sigma^{-1}(y(s))b(y(s), \alpha)|^2 ds.$$

Given $\varepsilon > 0$, there are $T_2 > T_1 > T_0 > 0$ such that for all $\alpha \in A_0$ and $x \in \Gamma$,

$$(4.2) \quad P_x^\alpha\{\tau \geq T_2\} \leq \varepsilon, \quad P_x^\alpha\{\tau \leq T_1\} \leq \varepsilon,$$

$$(4.3) \quad E_x^\alpha \exp \xi_0^\alpha(T_0, T_1) = 1,$$

$$(4.4) \quad \sup_{x \in \Gamma, \alpha \in A_0} E_x^\alpha \exp 2\xi_0^\alpha(0, T_1) \leq K_1^2 < \infty,$$

$$(4.5) \quad (E|\exp \xi_0^\alpha(0, T_0) - 1|^2)^{1/2} \leq \varepsilon.$$

Let $\tau_{12} = (\tau \wedge T_2) \vee T_1$. By (4.2), we have

$$|E_x^\alpha f(\tilde{X}_1) - E_x^\alpha f(x(\tau_{12}))| \leq 4\varepsilon\|f\|.$$

Write

$$E_x^\alpha f(x(\tau_{12})) = E_x^\alpha E_{x(T_1)}^\alpha f(x(\tau_{12})) = E_x^\alpha f_1(x(T_1)),$$

where f_1 is defined in the obvious way and $\|f_1\| \leq \|f\|$. By use of a Girsanov measure transformation, (4.4), (4.5), and Schwarz's inequality, we can write

$$\begin{aligned} E_x^\alpha f_1(x(T_1)) &= E_x^\alpha f_1(y(T_1)) \exp \xi_0^\alpha(0, T_1) \\ &= E_x^\alpha E_{y(T_0)}^\alpha f_1(y(T_1)) \exp \xi_0^\alpha(T_0, T_1) + \varepsilon' \\ &= E_x^\alpha f_2(y(T_0)) + \varepsilon', \end{aligned}$$

where $\|\varepsilon'\| \leq \varepsilon K_1 \|f\|$, f_2 is defined in the obvious way, and $\|f_2\| \leq \|f\|$. Note that f_2 depends on α , but $y(T_0)$ does not.

By the above estimates and arbitrariness of ε , we need only show the equicontinuity of the set $\{E_x^\alpha f_2(y(T_0)) : \|f_2\| \leq 1, \alpha \in A_0, f_2 \in L^\infty(\Gamma)\}$. Since $y(T_0)$ has a bounded density with respect to Lebesgue measure, using characteristic functions, we can write

$$E_x^\alpha f_2(y(T_0)) = \frac{1}{(2\pi)^r} \int f_2(y) dy \left\{ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (\exp -iu'y) E_x^\alpha \exp iu'y(T_0) du \right\}.$$

We have

$$|E_x^\alpha \exp iu'y(T_0)| \leq \exp -O(|u|^2),$$

where $O(\cdot)$ can be chosen independently of $x \in \Gamma$ and $\alpha \in A_0$. Also, the bracketed term is the density (modulo a proportionality factor $(1/(2\pi)^r)$) of $y(T_0)$ and is bounded by $\exp -O(|y|^2)$, where $O(\cdot)$ can be chosen independently of $x \in \Gamma$ and $\alpha \in A_0$. Thus, we need only prove that $E_x^\alpha \exp iu'y(T_0)$ is x -continuous on each bounded u -set. This follows from the Feller property of $y(\cdot)$. \square

COROLLARY. *Suppose that Assumptions 2.1–2.4 hold. Then the transition function $\tilde{P}(x, n, B | \alpha) \equiv E_x^\alpha I_B(\tilde{X}_n)$ is continuous in x , uniformly in B , n , and $\alpha \in A_0$. Also $\tilde{\mu}(B, \alpha_0 + \delta\alpha) \rightarrow \tilde{\mu}(B, \alpha_0)$, uniformly in $B \in \mathcal{B}(\Gamma)$.*

Proof. The first assertion is a direct consequence of the lemma. Let $g \in L^\infty(\Gamma)$, $\|g\| \leq 1$. Then, by the invariance of $\tilde{\mu}(\alpha)$,

$$\langle \tilde{\mu}(\alpha_0 + \delta\alpha), g \rangle = \int_{\Gamma} \tilde{\mu}(dx, \alpha_0 + \delta\alpha) E_x^{\alpha_0 + \delta\alpha} g(\tilde{X}_1).$$

A measure transformation argument and the continuity of $b(\cdot)$ can be used to show that, as $\delta\alpha \rightarrow 0$, $E_x^{\alpha_0 + \delta\alpha} g(\tilde{X}_1)$ converges to $E_x^{\alpha_0} g(\tilde{X}_1)$, uniformly in $x \in \Gamma$. The latter function is continuous on Γ by Lemma 4.1. In fact, the continuity and the convergence is uniform in g . From this, the invariance of $\tilde{\mu}(\alpha_0)$ and the weak convergence $\tilde{\mu}(\alpha_0 + \delta\alpha) \Rightarrow \tilde{\mu}(\alpha_0)$ (see Lemma 4.3 below), we have

$$\begin{aligned} \lim_{\delta\alpha \rightarrow 0} \langle \tilde{\mu}(\alpha_0 + \delta\alpha), g \rangle &= \int_{\Gamma} \tilde{\mu}(dx, \alpha_0) E_x^{\alpha_0} g(\tilde{X}_1) \\ &= \int_{\Gamma} \tilde{\mu}(dx, \alpha_0) g(x), \end{aligned}$$

where the convergence is uniform in g : $\|g\| \leq 1$. \square

The next lemma will be used to get the differentiability of $\mu(\alpha)$ at α_0 from that of $\tilde{\mu}(\alpha)$, via (3.3).

LEMMA 4.2. *Suppose that Assumptions 2.1–2.5 hold. Then for $f \in L^\infty(\Gamma)$, as $\delta\alpha \rightarrow 0$,*

$$[\tilde{P}(\alpha_0 + \delta\alpha) - \tilde{P}(\alpha_0)]f/\delta\alpha$$

converges (uniformly in x) to the function with values $E_x^{\alpha_0} f(\tilde{X}_1)Z(\tau, \alpha_0)$. The limit is continuous and the convergence is uniform for f : $\|f\| \leq 1$. The set

$$\{E_x^{\alpha_0} Z(\tau, \alpha_0) f(\tilde{X}_1), \|f\| \leq 1, f \in L^\infty(\Gamma), \alpha \in A_0\}$$

is equicontinuous. The same result holds for the convergence

$$\begin{aligned} &\frac{1}{\delta\alpha} \left[E_x^{\alpha_0 + \delta\alpha} \int_0^\tau f(x(s)) ds - E_x^{\alpha_0} \int_0^\tau f(x(s)) ds \right] \\ &\rightarrow E_x^{\alpha_0} \int_0^\tau f(x(s)) ds Z(\tau, \alpha_0) = E_x^{\alpha_0} \int_0^\tau f(x(s)) Z(s, \alpha_0) ds. \end{aligned}$$

Proof. The proof of the last assertion is very similar to that of the prior assertions and will be omitted. By an argument analogous to that of Lemma 4.1, we can prove the equicontinuity of the cited set of functions. We will prove only the first assertion of the lemma. For $T < \infty$, via a Girsanov measure transformation,

$$\begin{aligned} \frac{E_x^{\alpha_0 + \delta\alpha} f(x(\tau \wedge T)) - E_x^{\alpha_0} f(x(\tau \wedge T))}{\delta\alpha} &= E_x^{\alpha_0} f(x(\tau \wedge T)) [\exp \xi(0, T; \alpha_0, \delta\alpha) - 1] / \delta\alpha \\ (4.6) \qquad \qquad \qquad &= E_x^{\alpha_0} f(x(\tau \wedge T)) [\exp \xi(0, \tau \wedge T; \alpha_0, \delta\alpha) - 1] / \delta\alpha. \end{aligned}$$

We have, by Assumption 2.5 and Theorem 2.1,

$$\lim_{\delta\alpha \rightarrow 0} \lim_{T \rightarrow \infty} E_x^{\alpha_0} \left[\frac{\exp \xi(0, \tau \wedge T; \alpha_0, \delta\alpha) - 1}{\delta\alpha} - Z(\tau, \alpha_0) \right]^2 = 0,$$

where the limit is attained uniformly in $x \in \Gamma$. The first assertion of the lemma follows from this and (4.6). \square

The next corollary shows that the setwise derivative of $\tilde{\mu}(\alpha)$ at α_0 is absolutely continuous with respect to $\tilde{\mu}(\alpha_0)$.

COROLLARY. *Suppose that Assumptions 2.1–2.4 hold. Define the set function \tilde{v} by*

$$\tilde{v}(B) = \lim_{\delta\alpha \rightarrow 0} \frac{1}{\delta\alpha} \int_{\Gamma} \tilde{\mu}(dx, \alpha_0) [\tilde{P}(x, 1, B \mid \alpha_0 + \delta\alpha) - \tilde{P}(x, 1, B \mid \alpha_0)].$$

Then there is $G \in L^1(\tilde{\mu}(\alpha_0))$ such that $\langle \tilde{\mu}(\alpha_0), G \rangle = 0$ and

$$\tilde{v}(B) = \int_B \tilde{\mu}(dx, \alpha_0) G(x).$$

The limit is uniform in B .

Proof. By the lemma, the limit is

$$\int_{\Gamma} \tilde{\mu}(dx, \alpha_0) E_x^{\alpha_0} Z(\tau, \alpha_0) I_B(\tilde{X}_1),$$

and the limit is taken on uniformly in B . (In fact, $E_x^{\alpha_0} Z(\tau, \alpha_0) I_B(\tilde{X}_1)$ is continuous, uniformly in B .) Both $\tilde{\mu}(\alpha_0)$ and the measure defined by the limit are mutually absolutely continuous with respect to Lebesgue measure, since the transition probability $\tilde{P}(x, 1, \cdot \mid \alpha_0)$ is. Let G denote the Radon–Nikodym derivative of \tilde{v} with respect to $\tilde{\mu}(\alpha_0)$. Since $E_x^{\alpha_0} Z(\tau, \alpha_0) I_{R^r}(\tilde{X}_1) = 0$, we have $\langle \tilde{\mu}(\alpha_0), G \rangle = 0$. \square

LEMMA 4.3. *Suppose that Assumptions 2.1–2.4 hold. Then $\tilde{\mu}(\alpha_0 + \delta\alpha) \Rightarrow \tilde{\mu}(\alpha_0)$.*

Proof. The proof follows from the uniqueness of $\tilde{\mu}(\alpha_0)$ and the convergence $\tilde{P}(x, 1, B \mid \alpha_0 + \delta\alpha) \rightarrow \tilde{P}(x, 1, B \mid \alpha_0)$, uniformly for $x \in \Gamma$ (Lemma 4.1), and the details are omitted. \square

DEFINITION. Let $L_c^\infty(\Gamma) \subset L^\infty(\Gamma)$ be the “centered” subset for which $\langle \tilde{\mu}(\alpha_0), f \rangle = 0$. We identify functions in $L_c^\infty(\Gamma)$ that are equal almost everywhere (Lebesgue measure).

The following lemma is a key result for proving the differentiability of $\tilde{\mu}(\alpha)$ at α_0 . The representations that are used occur throughout the following.

LEMMA 4.4. *Suppose that Assumptions 2.1–2.4 hold. Then $(I - \tilde{P}(\alpha_0)): L_c^\infty(\Gamma) \rightarrow L_c^\infty(\Gamma)$ is invertible.*

Proof. The fact that $\tilde{P}(\alpha_0)$ maps $L_c^\infty(\Gamma)$ into $L_c^\infty(\Gamma)$ follows from the fact that $\tilde{\mu}(\alpha_0)$ is an invariant measure for the transition function $\tilde{P}(x, n, \cdot \mid \alpha_0)$. We prove the invertability by simply exhibiting the inverse. Let $f \in L_c^\infty(\Gamma)$. Then it is easily seen from (3.6) and the definition of $(I - \tilde{P}(\alpha_0))$ that the “inverse” defined by

$$(4.7) \quad (I - \tilde{P}(\alpha_0))^{-1} f(x) \equiv \sum_{n=0}^{\infty} \tilde{P}^n(\alpha_0) f(x) = \sum_{n=0}^{\infty} E_x^{\alpha_0} f(\tilde{X}_n)$$

satisfies our needs. \square

COROLLARY. Suppose that Assumptions 2.1–2.4 hold. Then $(I - \tilde{P}(\alpha_0)): C_c(\Gamma) \rightarrow C_c(\Gamma)$ is invertible.

Proof. By Lemma 4.1, $\tilde{P}(\alpha_0)C_c(\Gamma) \subset C_c(\Gamma)$. The rest of the proof is as for the lemma. \square

THEOREM 4.1. Suppose that Assumptions 2.1–2.5 hold. Then $\tilde{\mu}_\alpha(\alpha_0)$ exists in the sense of setwise convergence and satisfies, for $f \in L^\infty(\Gamma)$ and all n ,

$$(4.8) \quad \langle \tilde{\mu}_\alpha(\alpha_0), f \rangle = \langle \tilde{\mu}(\alpha_0), \tilde{P}_\alpha^n(\alpha_0)f \rangle + \langle \tilde{\mu}_\alpha(\alpha_0), \tilde{P}^n(\alpha_0)f \rangle,$$

where

$$\tilde{P}_\alpha^n(\alpha_0)f(x) = \frac{d}{d\alpha} E_x^\alpha f(\tilde{X}_n) \Big|_{\alpha_0}.$$

Proof. For $f \in L_c^\infty(\Gamma)$, we have

$$(4.9) \quad \begin{aligned} \langle \tilde{\mu}(\alpha) - \tilde{\mu}(\alpha_0), f \rangle &= \langle \tilde{\mu}(\alpha), \tilde{P}(\alpha)f \rangle - \langle \tilde{\mu}(\alpha_0), \tilde{P}(\alpha_0)f \rangle \\ &= \langle \tilde{\mu}(\alpha) - \tilde{\mu}(\alpha_0), \tilde{P}(\alpha_0)f \rangle + \langle \tilde{\mu}(\alpha_0), (\tilde{P}(\alpha) - \tilde{P}(\alpha_0))f \rangle \\ &\quad + \langle \tilde{\mu}(\alpha) - \tilde{\mu}(\alpha_0), (\tilde{P}(\alpha) - \tilde{P}(\alpha_0))f \rangle. \end{aligned}$$

Write $\delta\tilde{\mu}(\alpha) = \tilde{\mu}(\alpha) - \tilde{\mu}(\alpha_0)$ and $\delta\tilde{P}(\alpha) = \tilde{P}(\alpha) - \tilde{P}(\alpha_0)$. Then (4.9) yields

$$(4.10) \quad \langle \delta\tilde{\mu}(\alpha)/\delta\alpha, (I - \tilde{P}(\alpha_0))f \rangle = \langle \tilde{\mu}(\alpha_0), \frac{\delta\tilde{P}(\alpha)}{\delta\alpha}f \rangle + \langle \delta\tilde{\mu}(\alpha), \frac{\delta\tilde{P}(\alpha)}{\delta\alpha}f \rangle.$$

By Lemma 4.2 and either Lemma 4.3 or the corollary to Lemma 4.1, the second right-hand term in (4.10) goes to zero as $\delta\alpha \rightarrow 0$ (uniformly in $f: \|f\| \leq 1$).

For $g \in L_c^\infty(\Gamma)$, define (use Lemma 4.4), $f = (I - \tilde{P}(\alpha_0))^{-1}g$. By Lemmas 4.2 and 4.4,

$$\frac{\delta\tilde{P}(\alpha)}{\delta\alpha} (I - \tilde{P}(\alpha_0))^{-1}g$$

converges (uniformly in x) to the function with values

$$E_x^{\alpha_0} f(\tilde{X}_1)Z(\tau, \alpha_0) = E_x^{\alpha_0} [Z(\tau, \alpha_0) \sum_{n=0}^{\infty} E_y^{\alpha_0} g(\tilde{X}_n) \Big|_{y=\tilde{X}_1}] \equiv \tilde{g}(x),$$

which is in $C_c(\Gamma)$. Hence

$$(4.11) \quad \lim_{\delta\alpha \rightarrow 0} \langle \delta\tilde{\mu}(\alpha)/\delta\alpha, g \rangle = \langle \tilde{\mu}(\alpha_0), \tilde{g} \rangle.$$

Since $g \in L_c^\infty(\Gamma)$, and $L_c^\infty(\Gamma) = L^\infty(\Gamma)$ modulo constant functions, (4.11) gives the desired setwise convergence.

The formula (4.8) follows in a similar way. \square

COROLLARY. Suppose that Assumptions 2.1–2.5 hold. Then $\tilde{\mu}_\alpha(\alpha_0)$ exists in the sense of weak convergence.

Remark. The corollary is obviously a special case of the theorem. It can be proved directly via the method of proof of the theorem, simply by replacing all $L_c^\infty(\Gamma)$ by $C_c(\Gamma)$. This remark will be useful when working with the approximations in §7, since there we work with weak convergence only.

Since the existence of $\tilde{\mu}_\alpha(\alpha_0)$ is now established, we can turn our attention to $\mu_\alpha(\alpha_0)$.

THEOREM 4.2. *Suppose that Assumptions 2.1–2.5 hold. Then $\mu_\alpha(\alpha_0)$ exists in the sense of setwise convergence, and for $f \in L^\infty(R^r)$,*

$$\begin{aligned}
 \langle \mu_\alpha(\alpha_0), f \rangle &= \frac{1}{\tilde{\mu}(R^r, \alpha_0)} \left[\int_{\Gamma} \tilde{\mu}(dx, \alpha_0) E_x^{\alpha_0} \int_0^\tau f(x(s)) Z(s, \alpha_0) ds \right. \\
 &\quad \left. + \int_{\Gamma} \tilde{\mu}_\alpha(dx, \alpha_0) E_x^{\alpha_0} \int_0^\tau f(x(s)) ds \right] \\
 &\quad - \frac{\langle \tilde{\mu}(\alpha_0), f \rangle}{(\tilde{\mu}(R^r, \alpha_0))^2} \left[\int_{\Gamma} \tilde{\mu}(dx, \alpha_0) E_x^{\alpha_0} \int_0^\tau Z(s, \alpha_0) ds + \int_{\Gamma} \tilde{\mu}_\alpha(dx, \alpha_0) E_0^\alpha \tau \right] \\
 (4.12) \quad &= \frac{d}{d\alpha} \left[\frac{\int_{\Gamma} \tilde{\mu}(dx, \alpha) E_x^\alpha \int_0^\tau f(x(s)) ds}{\int_{\Gamma} \tilde{\mu}(dx, \alpha) E_x^\alpha \tau} \right]_{\alpha_0}.
 \end{aligned}$$

Also, $\mu_\alpha(\alpha_0)$ is absolutely continuous with respect to Lebesgue measure and has finite variation.

Proof. Let $f \in L^\infty(R^r)$. Define $\delta\mu(\alpha) = \mu(\alpha_0 + \delta\alpha) - \mu(\alpha_0)$ and define $\delta\tilde{\mu}(\alpha)$ analogously. Define the operator $\hat{P}(\alpha)$ on $L^\infty(R^r)$ by $\hat{P}(\alpha)f = E_x^\alpha \int_0^\tau f(x(s)) ds$. Let \mathbf{e} denote the function that is identically unity. We need to show the differentiability of

$$\langle \tilde{\mu}(\alpha), \tilde{P}(\alpha)f \rangle / \langle \tilde{\mu}(\alpha), \tilde{P}(\alpha)\mathbf{e} \rangle = \langle \mu(\alpha), f \rangle.$$

It will be sufficient to show the differentiability of the numerator only. This will be the first bracketed term in (4.12). We can write

$$\begin{aligned}
 &\frac{1}{\delta\alpha} [\langle \tilde{\mu}(\alpha_0 + \delta\alpha), \hat{P}(\alpha_0 + \delta\alpha)f \rangle - \langle \tilde{\mu}(\alpha_0), \hat{P}(\alpha_0)f \rangle] \\
 &= \left\langle \frac{\delta\tilde{\mu}(\alpha)}{\delta\alpha}, \hat{P}(\alpha_0)f \right\rangle + \left\langle \tilde{\mu}(\alpha_0), \frac{(\hat{P}(\alpha_0 + \delta\alpha) - \hat{P}(\alpha_0))}{\delta\alpha} f \right\rangle \\
 &\quad + \left\langle \delta\tilde{\mu}(\alpha_0), \frac{(\hat{P}(\alpha_0 + \delta\alpha) - \hat{P}(\alpha_0))}{\delta\alpha} f \right\rangle.
 \end{aligned}$$

By Lemma 4.2, the second term on the right converges to the first term in the first bracket on the right-hand side of (4.12). The first term on the right converges to the second term in the first bracket on the right-hand side of (4.12) by Theorem 4.1 and the fact that $\hat{P}(\alpha_0)f \in C(\Gamma)$. Similarly, the last term on the right goes to zero. Representation (4.12) implies the absolute continuity assertion since it equals zero if $f = 0$ almost everywhere (Lebesgue measure). It also implies the finite variation. \square

Theorem 4.3 essentially says that the α -derivative of $E_x^\alpha f(x(t))$ equals that of $\langle \mu(\alpha), f \rangle$ for large t .

THEOREM 4.3. *Suppose that Assumptions 2.1–2.5 hold. Then for $f \in L^\infty(R^r)$,*

$$\begin{aligned}
 (4.13) \quad &\lim_{t \rightarrow \infty} \frac{d}{d\alpha} \int \mu(dx, \alpha) E_x^\alpha f(x(t)) \Big|_{\alpha=\alpha_0} = \frac{d}{d\alpha} \langle \mu(\alpha), f \rangle \Big|_{\alpha=\alpha_0} \\
 &= \lim_{t \rightarrow \infty} \int \mu(dx, \alpha_0) (E_x^{\alpha_0} f(x(t)))_\alpha,
 \end{aligned}$$

$$\begin{aligned}
 (4.14) \quad & \lim_{t \rightarrow \infty} \frac{d}{d\alpha} \int \mu(dx, \alpha) E_x^\alpha \frac{1}{t} \int_0^t f(x(s)) ds \Big|_{\alpha=\alpha_0} \\
 &= \lim_{t \rightarrow \infty} \int \mu(dx, \alpha_0) (E_x^{\alpha_0} f(x(t)))_\alpha,
 \end{aligned}$$

and the limits exist.

Proof. By the differentiability proved in Theorem 4.2 we can write

$$\begin{aligned}
 (4.15) \quad & \frac{d}{d\alpha} \int \mu(dx, \alpha) f(x) = \frac{d}{d\alpha} \int \mu(dx, \alpha) E_x^\alpha f(x(t)) \Big|_{\alpha=\alpha_0} \\
 &= \int \mu_\alpha(dx, \alpha_0) E_x^{\alpha_0} f(x(t)) + \int \mu(dx, \alpha_0) (E_x^{\alpha_0} f(x(t)))_\alpha.
 \end{aligned}$$

As $t \rightarrow \infty$, $E_x^{\alpha_0} f(x(t)) \rightarrow \langle \mu(\alpha_0), f \rangle$ for $\mu(\alpha_0)$ -almost all x . Since $\mu_\alpha(\alpha_0)$ is absolutely continuous with respect to Lebesgue measure (Theorem 4.2), and $\mu(\alpha_0)$ and Lebesgue measure are mutually absolutely continuous, we have that $\mu_\alpha(\alpha_0)$ is absolutely continuous with respect to $\mu(\alpha_0)$. Also $\langle \mu_\alpha(\alpha_0), \text{constant function} \rangle = 0$. These facts imply that the first term on the right-hand side of (4.15) goes to zero as $t \rightarrow \infty$, which yields the assertion concerning (4.13). Expression (4.14) is proved in the same way. \square

5. A discrete time approximation. Since the paths of $x(\cdot)$ and $w(\cdot)$ are not physically available, we cannot evaluate (1.5) or use Theorem 4.2 or 4.3 as stated to get estimates of the derivatives $\langle \mu(\alpha_0), f \rangle_\alpha$ via the use of paths of $x(\cdot)$ or $w(\cdot)$. We must work with computable approximations to $x(\cdot)$ and $w(\cdot)$. In [1], two types of approximations were used for the finite time problem: the first was a discrete time approximation and the second, a Markov chain approximation. Each one has its own advantages, but simulation studies indicate that their overall numerical properties are similar. We will work with the discrete time approximation here. In this section, the approximation is defined. Some necessary stability results are proved in the next section. Among other things to be shown, the robustness properties of approximations to derivatives of invariant measures and ergodic costs will be clear.

For $\Delta > 0$ and $\delta w(n\Delta) = w(n\Delta + \Delta) - w(n\Delta)$, define $\{X_n^\Delta\}$ by $X_0^\Delta = x$ and

$$(5.1) \quad X_{n+1}^\Delta = X_n^\Delta + \Delta b(X_n^\Delta, \alpha_0) + \sigma(X_n^\Delta) \delta w(n\Delta).$$

Define the interpolation $x^\Delta(\cdot)$ to be the piecewise constant (on intervals $[n\Delta, n\Delta + \Delta)$) process with $x^\Delta(n\Delta) = X_n^\Delta$. Define $Z^\Delta(\cdot, \alpha_0)$ to be the piecewise constant (on intervals $[n\Delta, n\Delta + \Delta)$) process with value at $n\Delta$:

$$\begin{aligned}
 Z^\Delta(n\Delta, \alpha_0) &= \sum_{i=0}^{n-1} [\sigma^{-1}(X_i^\Delta, \alpha_0) b_\alpha(X_i^\Delta, \alpha_0)]' \delta w(i\Delta) \\
 &= \sum_{i=0}^{n-1} [b'_\alpha(X_i^\Delta, \alpha_0) a^{-1}(X_i^\Delta)] [\delta X_i^\Delta - \Delta b(X_i^\Delta, \alpha_0)],
 \end{aligned}$$

where $\delta X_i^\Delta \equiv X_{i+1}^\Delta - X_i^\Delta$. For $T = N/\Delta$, [1, §4] shows that

$$\begin{aligned}
 Q^\Delta(\alpha_0) &= \sum_{n=0}^{N-1} \Delta [k(X_n^\Delta, \alpha_0) Z^\Delta(n\Delta, \alpha_0) + k_\alpha(X_n^\Delta, \alpha_0)] \\
 &\quad + k_{0,\alpha}(X_N^\Delta, \alpha_0) + k_0(X_N^\Delta, \alpha_0) Z^\Delta(N\Delta, \alpha_0)
 \end{aligned}$$

or the centered form

$$\begin{aligned}\hat{Q}^\Delta(\alpha_0) = & \sum_{n=0}^{N-1} \Delta[(k(X_{n,\alpha_0}^\Delta) - E_x^{\alpha_0} k(X_n^\Delta, \alpha_0))Z^\Delta(n\Delta, \alpha_0) + k_\alpha(X_n^\Delta, \alpha_0)] \\ & + k_{0,\alpha}(X_N^\Delta, \alpha_0) + (k_0(X_N^\Delta, \alpha_0) - \bar{k}_0^{\Delta, \alpha_0})Z^\Delta(N\Delta, \alpha_0)\end{aligned}$$

are appropriate approximations to (1.5). The $\hat{Q}^\Delta(\alpha_0)$ will have the smaller variance.

In fact, we have

$$E_x^{\alpha_0} \hat{Q}^\Delta(\alpha_0) = E_x^{\alpha_0} Q^\Delta(\alpha_0) = \frac{d}{d\alpha} \left[E_x^\alpha \int_0^T k(x^\Delta(s), \alpha) ds + k_0(x^\Delta(T), \alpha) \right]_{\alpha=\alpha_0},$$

and we have the weak convergence

$$(Z^\Delta(\cdot, \alpha_0), Q^\Delta(\alpha_0), \hat{Q}^\Delta(\alpha_0), x^\Delta(\cdot)) \Rightarrow (Z(\cdot, \alpha_0), Q(\alpha_0), Q(\alpha_0), x(\cdot)).$$

We will obtain various “infinite time” extensions of this result in §7.

Analogous to the comment below (1.5'), to reduce computation while exploiting the (variance reduction) advantages of the centering, in the simulations we replace $E_x^{\alpha_0} k(X_n^\Delta, \alpha_0)$ by $E_x^{\alpha_0} k(X_N^\Delta, \alpha_0)$, with good results in general.

A remark on using the derivative estimators in recursive Monte Carlo optimization. Suppose that a sequence α_n^Δ is generated by a stochastic approximation scheme and converges to a minimizing $\bar{\alpha}^\Delta$. Then, via weak convergence methods, it can be shown that the limit of any convergent subsequence of $\bar{\alpha}^\Delta$ is optimal for the original problem, under appropriate regularity conditions.

6. Stability of the approximation. An analogue of Theorem 2.1 is needed for the $\{X_n^\Delta\}$ process. We will require the following additional condition.

Assumption 6.1.

(a) $V'_x(x)b(x, \alpha) \rightarrow -\infty$ as $|x| \rightarrow \infty$, uniformly for $\alpha \in A_0$.

(b) $\liminf_{|x| \rightarrow \infty} \inf_{\alpha \in A_0} \frac{|V'_x(x)b(x, \alpha)|}{|b(x, \alpha)|^2} > 0$.

THEOREM 6.1. *Suppose that Assumptions 2.1–2.3 and 6.1 hold. There is a compact set Q such that for each compact $Q_1 \supset Q$, we have for small $\rho > 0$, $\delta > 0$, and $\Delta < \delta$,*

$$(6.1) \quad \sup_{\alpha \in A_0} \sup_{x \in Q_1 - Q} E_x^\alpha \exp \rho \tau^\Delta < \infty,$$

where $\tau^\Delta = \min\{t: x^\Delta(t) \in Q\}$.

Proof. Let $X_0^\Delta = x$. For some $K_0 < \infty$, we have

$$\begin{aligned}A & \equiv E_x^\alpha \exp \rho[V(X_1^\Delta) - V(x)] \\ & \leq E_x^\alpha \exp \rho[V'_x(x)(b(x, \alpha)\Delta + \sigma(x)\delta w) + K_0(|\delta w|^2 + |b(x, \alpha)|^2\Delta^2)].\end{aligned}$$

Note that for $2rk\Delta < 1$,

$$E_x^\alpha \exp k|\delta w|^2 \leq 1/(1 - 2rk\Delta).$$

Thus, for small ρ , Δ , and $k_i > 0$, $1/k_1 + 1/k_2 = 1$, Hölder's inequality yields

$$\begin{aligned}& E_x^\alpha \exp \rho[V'_x(x)\sigma(x)\delta w + K_0|\delta w|^2] \\ & \leq [\exp k_1^2 \rho^2 \Delta V'_x(x)a(x)V_x(x)/2]^{1/k_1} \cdot \frac{1}{(1 - 2r\rho k_2 K_0 \Delta)^{1/k_2}}.\end{aligned}$$

Thus, for small ρ , Δ , and k_1 fixed near unity,

$$A \leq \exp \rho [V'_x(x)b(x, \alpha)\Delta + K_0|b(x, \alpha)|^2\Delta^2 \\ + \frac{k_1}{2}\rho\Delta V'_x(x)a(x)V_x(x) + 4rK_0\Delta].$$

Thus, there is a compact set Q and $\varepsilon_1 > 0$ such that for small ρ and for $x \notin Q$, $A \leq \exp -2\rho\varepsilon_1\Delta$. Thus for small ρ and $x \notin Q$,

$$E_x^\alpha \exp \rho\Delta\varepsilon_1 \cdot \exp \rho V(X_1^\Delta) \leq \exp \rho V(x).$$

Hence

$$E_x^\alpha \exp \rho\varepsilon_1\tau^\Delta \cdot \exp \rho V(x^\Delta(\tau^\Delta)) \leq \exp \rho V(x),$$

which yields the result, as in Theorem 2.1. \square

7. Ergodic properties of $\{X_n^\Delta\}$. We now set up the machinery so that results analogous to those in §§4 and 5 and the limits as $\Delta \rightarrow 0$ can be obtained. Define Γ , G , Γ_1 , and G_1 as in §3. Define the following stopping times:

$$\begin{aligned} \tau^{\Delta'} &= \inf\{t: x^\Delta(t) \notin G_1 - \Gamma_1\}, \\ \tau_1^\Delta &= \inf\{t: x^\Delta(t) \in G\}, \\ \tau_1^{\Delta'} &= \inf\{t > \tau_1^\Delta: x^\Delta(t) \notin G_1 - \Gamma_1\}. \end{aligned}$$

For $n > 1$,

$$\tau_n^\Delta = \inf\{t > \tau_{n-1}^{\Delta'}: x^\Delta(t) \in G\}, \quad \tau_n^{\Delta'} = \inf\{t > \tau_n^\Delta: x^\Delta(t) \notin G_1 - \Gamma_1\}.$$

For $x = x^\Delta(0) \in G$, we use τ^Δ to denote $\tau_2^\Delta - \tau_1^\Delta = \tau_2^\Delta$, the canonical return time to G .

By Theorem 6.1, there are G, G_1 such that (e.g., let G equal the set Q of Theorem 6.1)

$$(7.1) \quad \sup_{x \in G, \alpha \in A_0} E_x^\alpha \tau^\Delta < \infty, \quad \sup_{x \in G, \alpha \in A_0} E_x^\alpha \exp \rho \tau^\Delta < \infty,$$

for small ρ . Define $\tilde{X}_n^\Delta = x^\Delta(\tau_n^\Delta)$. For $\alpha \in A_0$, the process $\{\tilde{X}_n^\Delta, n \geq 0\}$ is a homogeneous, positive recurrent Markov chain with state space G . Let $\tilde{P}^\Delta(x, n, \cdot | \alpha)$ denote the transition function. There is a unique invariant measure $\tilde{\mu}^\Delta(\alpha)$. Analogously to the situation in §3, define the following:

$$\begin{aligned} \tau^\Delta(A) &= \int_0^{\tau^\Delta} I_A(x^\Delta(s))ds, \quad A = \text{Borel set in } R^r, \\ \bar{\mu}^\Delta(A, \alpha) &= \int_G \tilde{\mu}^\Delta(dx, \alpha) E_x^\alpha \tau^\Delta(A), \\ \mu^\Delta(A, \alpha) &= \bar{\mu}^\Delta(A, \alpha) / \bar{\mu}^\Delta(R^r, \alpha). \end{aligned}$$

The same argument used to show that $\mu(\alpha)$ is invariant for $x(\cdot)$ [2, p. 183] can be used to show that $\mu^\Delta(\alpha)$ is invariant for $\{X_n^\Delta\}$, under parameter α . We can now write for bounded measurable f :

$$(7.2) \quad \langle \mu^\Delta(\alpha), f \rangle = \frac{\int_G \tilde{\mu}^\Delta(dx, \alpha) E_x^\alpha \int_0^{\tau^\Delta} f(x^\Delta(s))ds}{\int_G \tilde{\mu}^\Delta(dx, \alpha) E_x^\alpha \tau^\Delta}.$$

Let $L^\infty(G)$ denote the set of bounded Borel measurable functions on G . Define the operator $\tilde{P}^\Delta(\alpha)$ on $L^\infty(G)$ by $\tilde{P}^\Delta(\alpha)f(x) = E_x^\alpha f(\tilde{X}_1^\Delta)$, $x \in G$.

LEMMA 7.1. *Suppose that Assumptions 2.1–2.4 and 6.1 hold. Then the set $\{\tilde{P}^\Delta(\alpha)L^\infty(G)$ (restricted to $\|f\| \leq 1$), $\Delta > 0$, $\alpha \in A_0\}$ is equicontinuous.*

Remark on the proof. Define the process $\bar{y}^\Delta(\cdot)$ to be the piecewise constant interpolation (intervals $[n\Delta, n\Delta + \Delta)$) of the process defined by $\bar{Y}_0^\Delta = x$, $\bar{Y}_{n+1}^\Delta = \bar{Y}_n^\Delta + \sigma(\bar{Y}_n^\Delta)\delta w(n\Delta)$. Then $\bar{y}^\Delta(\cdot) \Rightarrow y(\cdot)$, defined in Lemma 4.1. Define the Radon–Nikodym derivative $\exp \xi_0^{\alpha, \Delta}(0, T)$, where

$$\begin{aligned} \xi_0^{\alpha, \Delta}(0, T) &= \sum_{n=0}^{T/\Delta-1} [\sigma^{-1}(\bar{Y}_n^\Delta)b(\bar{Y}_n^\Delta, \alpha)]' \delta w(n\Delta) \\ &\quad - \frac{1}{2} \sum_{n=0}^{T/\Delta-1} |\sigma^{-1}(\bar{Y}_n^\Delta)b(\bar{Y}_n^\Delta, \alpha)|^2 \Delta. \end{aligned}$$

From this point on, the proof is nearly identical to that of Lemma 4.1 and is omitted.

THEOREM 7.1. *Suppose that Assumptions 2.1–2.4 and 6.1 hold, and let $\alpha = \alpha_0$. Then $\tilde{X}_k^\Delta \Rightarrow \tilde{X}_k$ if $\tilde{X}_0^\Delta \Rightarrow \tilde{X}_0$, and $\tilde{\mu}^\Delta(\alpha_0) \Rightarrow \tilde{\mu}(\alpha_0)$. In addition, $E_x^{\alpha_0} f(\tilde{X}_k^\Delta) \xrightarrow{\Delta} E_x^{\alpha_0} f(\tilde{X}_k)$ uniformly in $x \in G$ and in f in any equicontinuous set with $\|f\| \leq 1$. Also, $\tilde{\mu}^\Delta(\alpha_0 + \delta\alpha) \Rightarrow \tilde{\mu}^\Delta(\alpha_0)$ as $\delta\alpha \rightarrow 0$ and $\mu^\Delta(\alpha_0) \Rightarrow \mu(\alpha_0)$. Finally,*

$$\bar{f}^{\Delta, \alpha_0} \equiv \langle \mu^\Delta(\alpha_0), f \rangle \rightarrow \langle \mu(\alpha_0), f \rangle \equiv \bar{f}^{\alpha_0},$$

uniformly for f in any equicontinuous set with $\|f\| \leq 1$.

Proof. Note that $(x^\Delta(\cdot), \tau^\Delta) \rightarrow (x(\cdot), \tau)$ uniformly in $x \in G$ in the sense that $E_x^{\alpha_0} F(x^\Delta(\cdot), \tau^\Delta) \rightarrow E_x^{\alpha_0} F(x(\cdot), \tau)$ uniformly in $x \in G$, for any bounded and continuous real valued $F(\cdot)$. The weak convergence $\tilde{X}_k^\Delta \Rightarrow \tilde{X}_k$ (if $\tilde{X}_0^\Delta \Rightarrow \tilde{X}_0$) follows from the uniform integrability of $\{\tau^\Delta, \Delta > 0, \alpha \in A_0\}$ and the (uniform) weak convergence of $x^\Delta(\cdot)$ to $x(\cdot)$. The asserted weak convergence can be proved by a standard martingale method [5], [6] (and using the nondegeneracy of $a(\cdot)$ and the smoothness of Γ, Γ_1 to get the weak convergence of τ^Δ). In fact, a standard weak convergence method can be used to get $\tilde{P}^\Delta(\alpha_0)f \rightarrow \tilde{P}(\alpha_0)f$, uniformly in f in any equicontinuous set in $C(G)$.

Now, for $f \in C(G)$, by the invariance of $\mu^\Delta(\alpha_0)$, we can write

$$\langle \tilde{\mu}^\Delta(\alpha_0), f \rangle = \langle \tilde{\mu}^\Delta(\alpha_0), \tilde{P}^\Delta(\alpha_0)f \rangle \equiv \bar{f}^{\Delta, \alpha_0}.$$

$\{\tilde{\mu}^\Delta(\alpha_0), \Delta > 0\}$ is obviously tight since G is compact. If $\hat{\mu}(\alpha_0)$ is the limit of a weakly convergent subsequence, then by the last expression, we have

$$\langle \hat{\mu}(\alpha_0), f \rangle = \langle \hat{\mu}(\alpha_0), \tilde{P}(\alpha_0)f \rangle, \quad f \in C(G),$$

which yields $\hat{\mu}(\alpha_0) = \tilde{\mu}(\alpha_0)$.

Now use (7.2), the weak convergence $\{\tau^\Delta, x^\Delta(\cdot)\} \Rightarrow \{\tau, x(\cdot)\}$, and the uniform integrability of $\{\tau^\Delta\}$ (Theorem 6.1) and $\tilde{\mu}^\Delta(\alpha_0) \Rightarrow \tilde{\mu}(\alpha_0)$ to get $\mu^\Delta(\alpha_0) \Rightarrow \mu(\alpha_0)$. The last assertion of the theorem is also proved by an argument by contradiction, and the proof is omitted. \square

An analogue of (3.6). The following lemma is needed to get an analogue of Lemma 4.4.

LEMMA 7.2. *Suppose that Assumptions 2.1–2.3 and 6.1 hold. Let k be such that $C\gamma^k \equiv \lambda < 1$ (see (3.5)). Let $C'(G) \subset C(G)$ be an equicontinuous set. Then*

$$(7.3) \quad \overline{\lim}_{\Delta \rightarrow 0} \sup_{f \in C'(G)} \frac{|E_x^{\alpha_0} f(\tilde{X}_k^\Delta) - \bar{f}^{\Delta, \alpha_0}|}{\|f - \bar{f}^{\Delta, \alpha_0}\|} \leq \lambda.$$

Equivalently, there are $\psi_1 < 1$, $C_1 < \infty$, such that for small $\Delta > 0$,

$$(7.4) \quad \|\tilde{P}^\Delta(\alpha_0)^n f - \tilde{f}^{\Delta, \alpha_0}\| \leq C_1 \psi_1^n \|f - \tilde{f}^{\Delta, \alpha_0}\|.$$

Proof. Suppose that (7.3) is false. Then there is $x_n \rightarrow x \in G$, $\Delta_n \rightarrow 0$, $\lambda_n \geq \lambda_0 > \lambda$, $f_n \in C'(G)$, $f_n \rightarrow f \in C'(G)$, such that

$$\frac{|E_{x_n}^{\alpha_0} f_n(\tilde{X}_k^{\Delta_n}) - \tilde{f}_n^{\Delta_n, \alpha_0}|}{\|f_n - \tilde{f}_n^{\Delta_n, \alpha_0}\|} \geq \lambda_n.$$

Without loss of generality, we can suppose that the infima of the denominators are positive. Then we can write

$$(7.5) \quad \frac{|E_x^{\alpha_0} f(\tilde{X}_k) - \tilde{f}^{\alpha_0}|}{\|f - \tilde{f}^{\Delta_n, \alpha_0}\|} \geq \frac{|E_{x_n}^{\alpha_0} f_n(\tilde{X}_k^{\Delta_n}) - \tilde{f}_n^{\Delta_n, \alpha_0}|}{\|f_n - \tilde{f}_n^{\Delta_n, \alpha_0}\|} \frac{\|f - \tilde{f}_n^{\Delta_n, \alpha_0}\|}{\|f_n - \tilde{f}_n^{\Delta_n, \alpha_0}\|}.$$

The last two terms on the right go to zero by the weak convergence $\tilde{X}_k^{\Delta_n} \Rightarrow \tilde{X}_k$ (initial conditions $\tilde{X}_0^\Delta = x_n$, $\tilde{X}_0 = x$, respectively), $\tilde{\mu}^{\Delta_n}(\alpha_0) \Rightarrow \tilde{\mu}(\alpha_0)$, and the convergence $f_n \rightarrow f$. The left side of (7.5) goes to $|E_x^{\alpha_0} f(\tilde{X}_k) - f^{\alpha_0}|/\|f - \tilde{f}^{\alpha_0}\| \leq C\psi^k = \lambda$, and we have a contradiction.

Inequality (7.4) follows from (7.3) by letting $\psi_1^k = (\lambda + \delta\lambda)$ for small $\delta\lambda > 0$, and iterating. \square

LEMMA 7.3. *Suppose that Assumptions 2.1–2.5 and 6.1 hold. Then for $f \in L^\infty(G)$,*

$$\frac{[\tilde{P}^\Delta(\alpha_0 + \delta\alpha) - \tilde{P}^\Delta(\alpha_0)]f}{\delta\alpha}$$

converges (as $\delta\alpha \rightarrow 0$) to the function $\tilde{P}_\alpha^\Delta(\alpha_0)f$ with values

$$E_x^{\alpha_0} Z^\Delta(\tau^\Delta, \alpha_0) f(\tilde{X}_1^\Delta) = \frac{d}{d\alpha} E_x^\alpha f(\tilde{X}_1^\Delta) \Big|_{\alpha=\alpha_0}.$$

The limit is continuous and the convergence is uniform in Δ , $x \in G$, and in $f \in C(G)$ for $\|f\| \leq 1$. The set $\{E_x^{\alpha_0} Z^\Delta(\tau^\Delta, \alpha_0) f(\tilde{X}_1^\Delta), \Delta > 0, f \in C(G), \|f\| \leq 1\}$ is equicontinuous.

The same result holds for the convergence

$$\begin{aligned} & \frac{1}{\delta\alpha} \left[E_x^{\alpha_0 + \delta\alpha} \int_0^{\tau^\Delta} f(x^\Delta(s)) ds - E_x^{\alpha_0} \int_0^{\tau^\Delta} f(x^\Delta(s)) ds \right] \\ & \rightarrow E_x^{\alpha_0} \int_0^{\tau^\Delta} f(x^\Delta(s)) Z^\Delta(s, \alpha_0) ds. \end{aligned}$$

The proof is analogous to that of Lemma 4.2 but uses the Radon–Nikodym derivative introduced in the remark under Lemma 7.1, and is omitted.

THEOREM 7.2. *Suppose that Assumptions 2.1–2.5 and 6.1 hold. Then $\tilde{\mu}_\alpha^\Delta(\alpha_0)$ and $\mu_\alpha^\Delta(\alpha_0)$ exist in the sense of weak convergence.*

Proof. Let $C_c^\Delta(G)$ be the subset of $C(G)$ for which $\langle f, \tilde{\mu}^\Delta(\alpha_0) \rangle = 0$. Following the proof of Lemma 4.4 and its corollary, we first show the invertability of $(I - \tilde{P}^\Delta(\alpha_0))$ on $C_c(G)$, on which we identify functions that are equal almost everywhere ($\tilde{\mu}^\Delta(\alpha_0)$). By Lemma 7.1 and the fact that $\tilde{\mu}^\Delta(\alpha_0)$ is an invariant measure for the transition function that defines $\tilde{P}^\Delta(\alpha_0)$, for $f \in C_c^\Delta(G)$ the sum below converges, and we have $(I - \tilde{P}^\Delta(\alpha_0))C_c^\Delta(G) \subset C_c^\Delta(G)$. By Lemma 7.2, we obviously have

$$(I - \tilde{P}^\Delta(\alpha_0)) \sum_{n=0}^{\infty} (\tilde{P}^\Delta(\alpha_0))^n f = \sum_{n=0}^{\infty} (\tilde{P}^\Delta(\alpha_0))^n (I - \tilde{P}^\Delta(\alpha_0)) f = f.$$

These facts yield that the inverse is

$$(7.6) \quad g^\Delta = (I - \tilde{P}^\Delta(\alpha_0))^{-1} f = \sum_{n=0}^{\infty} (\tilde{P}^\Delta(\alpha_0))^n f.$$

By Lemmas 7.1 and 7.2, the sum on the right side converges uniformly in Δ , and it is equicontinuous for $f \in C_c^\Delta(G)$, $\|f\| \leq 1$, $\Delta > 0$.

We can now use a proof analogous to that of Theorem 4.1 (but using weak rather than setwise convergence) together with Lemma 7.3 and the weak convergence $\tilde{\mu}^\Delta(\alpha_0 + \delta\alpha) \Rightarrow \tilde{\mu}^\Delta(\alpha_0)$ to get the existence of $\tilde{\mu}_\alpha^\Delta(\alpha_0)$ in the sense of weak convergence, and the few details are omitted. To get the existence of $\mu_\alpha^\Delta(\alpha_0)$ in the sense of weak convergence, use representation (7.2) and the α -differentiability of $\tilde{\mu}^\Delta(\alpha)$, $E_x^\alpha \int_0^{\tau^\Delta} f(x^\Delta(s)) ds$ at $\alpha = \alpha_0$, and $E_{x_0}^\alpha \tau^\Delta > 0$. The details are like those of Theorem 4.2, but the proof uses the equicontinuity of $\{E_x^\alpha \int_0^{\tau^\Delta} f(x^\Delta(s)) ds\}$ (in $f \in C(G)$, $\Delta > 0$, $\|f\| \leq 1$, $\alpha \in A_0$), the weak convergence, and the uniform integrability of $\{\tau^\Delta, \text{small } \Delta > 0, \alpha \in A_0\}$. \square

COROLLARY. *Assume the conditions of Theorem 7.2 hold. Then $\tilde{\mu}_\alpha^\Delta(\alpha_0)$ exists in the sense of setwise convergence. Also, $\{\tilde{\mu}_\alpha^\Delta(\alpha_0), \text{small } \Delta > 0\}$ is of bounded variation. For $g \in L^\infty(G)$, there is a unique $f \in L^\infty(G)$ such that*

$$(I - \tilde{P}^\Delta(\alpha_0))f = g - \tilde{g}^{\Delta, \alpha_0}$$

and $\tilde{f}^{\Delta, \alpha_0} = 0$.

Proof. Let $f \in L^\infty(G)$. Analogous to (4.9), write $\delta\tilde{P}(\alpha) = \tilde{P}^\Delta(\alpha_0 + \delta\alpha) - \tilde{P}^\Delta(\alpha_0)$, $\delta\tilde{\mu}^\Delta(\alpha) = \tilde{\mu}^\Delta(\alpha_0 + \delta\alpha) - \tilde{\mu}^\Delta(\alpha_0)$, and

$$(7.7) \quad \begin{aligned} \left\langle \frac{\delta\tilde{\mu}^\Delta(\alpha)}{\delta\alpha}, f \right\rangle &= \left\langle \frac{\delta\tilde{\mu}^\Delta(\alpha)}{\delta\alpha}, \tilde{P}^\Delta(\alpha_0)f \right\rangle \\ &+ \left\langle \tilde{\mu}^\Delta(\alpha_0), \frac{\delta\tilde{P}^\Delta(\alpha)}{\delta\alpha} f \right\rangle + \left\langle \delta\tilde{\mu}^\Delta(\alpha), \frac{\delta\tilde{P}^\Delta(\alpha_0)}{\delta\alpha} f \right\rangle \end{aligned}$$

By Lemma 7.3, $(\delta\tilde{P}^\Delta(\alpha)/\delta\alpha)f$ converges to a continuous function, uniformly in $x \in G$. This and $\delta\tilde{\mu}^\Delta(\alpha) \Rightarrow$ zero measure imply that the last term on the right of (7.7) tends to zero, as $\delta\alpha \rightarrow 0$. Furthermore, since $\tilde{\mu}_\alpha^\Delta(\alpha_0)$ exists in the sense of weak convergence by the theorem. The second term on the right of (7.7) tends to $\langle \tilde{\mu}^\Delta(\alpha_0), \tilde{P}_\alpha^\Delta(\alpha_0)f \rangle$.

Since $\tilde{P}^\Delta(\alpha_0)f$ is continuous (Lemma 7.1), and $\tilde{\mu}_\alpha^\Delta(\alpha_0)$ exists in the sense of weak convergence, the first term on the right tends to $\langle \tilde{\mu}_\alpha^\Delta(\alpha_0), \tilde{P}^\Delta(\alpha_0)f \rangle$. Thus the limit

of the left side of (7.7) exists. Now, the form of the limit of the right side implies that $\tilde{\mu}_\alpha^\Delta(\alpha_0)$ exists in the sense of setwise convergence.

Rewrite (7.7) as

$$\left\langle \tilde{\mu}_\alpha^\Delta(\alpha_0), (I - \tilde{P}^\Delta(\alpha_0))f \right\rangle = \left\langle \tilde{\mu}^\Delta(\alpha_0), \tilde{P}^\Delta(\alpha_0)f \right\rangle.$$

For $g \in L^\infty(G)$, set $\tilde{g} = g - \tilde{g}^{\Delta, \alpha_0}$ and define

$$(7.8) \quad \begin{aligned} f^\Delta &= \sum_{n=0}^{\infty} (\tilde{P}^\Delta(\alpha_0))^n \tilde{g} \\ &= \tilde{g} + \sum_{n=0}^{\infty} (\tilde{P}^\Delta(\alpha_0))^n (\tilde{P}^\Delta(\alpha_0)\tilde{g}). \end{aligned}$$

The sum converges uniformly in g, Δ , for $\|g\| \leq 1$, since $\{\tilde{P}^\Delta(\alpha_0)g, \Delta > 0, g \in L^\infty(G), \|g\| \leq 1\}$ is equicontinuous by Lemma 7.1. The uniqueness assertion follows.

Thus $(I - \tilde{P}^\Delta(\alpha_0))f^\Delta = \tilde{g}$ and

$$\langle \tilde{\mu}_\alpha^\Delta(\alpha_0), \tilde{g} \rangle = \langle \tilde{\mu}_\alpha^\Delta(\alpha_0), g \rangle = \langle \tilde{\mu}^\Delta(\alpha_0), \tilde{P}^\Delta(\alpha_0)f^\Delta \rangle.$$

The bounded variation assertion follows from this representation. \square

The following is the convergence theorem for the discretizations.

THEOREM 7.3. *Suppose that Assumptions 2.1–2.5 and 6.1 hold. Then $\tilde{\mu}_\alpha^\Delta(\alpha_0)$ converges setwise to $\tilde{\mu}_\alpha(\alpha_0)$ and $\mu_\alpha^\Delta(\alpha_0)$ converges setwise to $\mu_\alpha(\alpha_0)$.*

Proof. Let $f \in L^\infty(G)$. Let g^Δ and g , respectively, be the unique solutions in $L^\infty(G)$ (Theorem 7.2, Lemma 4.4) to

$$(I - \tilde{P}^\Delta(\alpha_0))g^\Delta = f - \tilde{f}^{\Delta, \alpha_0}, \quad (I - \tilde{P}(\alpha_0))g = f - \tilde{f}^{\alpha_0}.$$

Note that

$$\begin{aligned} \frac{d}{d\alpha} \langle \tilde{\mu}^\Delta(\alpha), g^\Delta \rangle|_{\alpha=\alpha_0} &= \langle \tilde{\mu}_\alpha^\Delta(\alpha_0), g^\Delta \rangle \\ &= \langle \tilde{\mu}_\alpha^\Delta(\alpha_0), \tilde{P}^\Delta(\alpha_0)g^\Delta \rangle + \langle \tilde{\mu}^\Delta(\alpha_0), \tilde{P}_\alpha^\Delta(\alpha_0)g^\Delta \rangle. \end{aligned}$$

Then we can write

$$(7.9) \quad \begin{aligned} \langle \tilde{\mu}_\alpha^\Delta(\alpha_0), f \rangle &= \langle \tilde{\mu}_\alpha^\Delta(\alpha_0), f - \tilde{f}^{\Delta, \alpha_0} \rangle \\ &= \langle \tilde{\mu}_\alpha^\Delta(\alpha_0), (I - \tilde{P}^\Delta(\alpha_0))g^\Delta \rangle \\ &= \int \tilde{\mu}^\Delta(dx, \alpha_0) \frac{d}{d\alpha} E_x^\alpha g^\Delta(\tilde{X}_1^\Delta)|_{\alpha=\alpha_0}. \end{aligned}$$

We have

$$\frac{d}{d\alpha} E_x^\alpha g^\Delta(\tilde{X}_1^\Delta)|_{\alpha=\alpha_0} = E_x^{\alpha_0} g^\Delta(\tilde{X}_1^\Delta) Z^\Delta(\tau^\Delta, \alpha_0).$$

Now, note that the sum in (7.6) converges uniformly in Δ (Lemma 7.2); hence $g^\Delta \rightarrow g$, since $\tilde{P}^\Delta(\alpha_0)^n f \rightarrow \tilde{P}(\alpha_0)^n f$. Using this, the weak convergence of $\{\tilde{X}_1^\Delta, Z^\Delta(\tau^\Delta, \alpha_0)\}$, the uniform integrability of $\{Z^\Delta(\tau^\Delta, \alpha_0), \Delta > 0\}$, the fact that the functions on the right side of (7.9) converge uniformly in $x \in G$ to the continuous limit, and the fact that $\tilde{\mu}^\Delta(\alpha_0) \Rightarrow \tilde{\mu}(\alpha_0)$ yields that the limit as $\Delta \rightarrow 0$ of the right side of (7.9) is

$$\int \tilde{\mu}(dx, \alpha_0) E_x^{\alpha_0} g(\tilde{X}_1) Z(\tau, \alpha_0) = \int \tilde{\mu}(dx, \alpha_0) \frac{d}{d\alpha} E_x^\alpha g(\tilde{X}_1)|_{\alpha=\alpha_0}$$

$$= \langle \tilde{\mu}_\alpha(\alpha_0), (I - \tilde{P}(\alpha_0))g \rangle = \langle \tilde{\mu}_\alpha(\alpha_0), f - \tilde{f}^{\alpha_0} \rangle = \langle \tilde{\mu}_\alpha(\alpha_0), f \rangle.$$

Thus $\langle \tilde{\mu}_\alpha^\Delta(\alpha_0), f \rangle \rightarrow \langle \tilde{\mu}_\alpha(\alpha_0), f \rangle$, which yields the setwise convergence of $\tilde{\mu}_\alpha^\Delta(\alpha_0)$ to $\tilde{\mu}_\alpha(\alpha_0)$. The setwise convergence of $\mu^\Delta(\alpha_0)$ to $\mu(\alpha_0)$ follows from representation (7.2). For example, to get the limit of the derivative of the denominator of (7.2), note that the derivative of the denominator is

$$\int_G \tilde{\mu}_\alpha^\Delta(dx, \alpha_0) E_x^{\alpha_0} \tau^\Delta + \int_G \tilde{\mu}_\alpha^\Delta(dx, \alpha_0) \frac{d}{d\alpha} E_x^\alpha \tau^\Delta \Big|_{\alpha=\alpha_0}.$$

Then use the representation

$$\frac{d}{d\alpha} E_x^\alpha \tau^\Delta \Big|_{\alpha=\alpha_0} = E_x^{\alpha_0} \tau^\Delta Z^\Delta(\tau^\Delta, \alpha_0),$$

and the proved convergence and uniform integrability (where appropriate) results for $\tilde{\mu}_\alpha^\Delta(\alpha_0)$, $X^\Delta(\cdot)$, τ^Δ , $Z^\Delta(\tau^\Delta, \alpha_0)$. \square

Below is a finite time approximation theorem, which shows that the derivative $\langle \mu_\alpha(\alpha_0), f \rangle$ of the ergodic cost can be arbitrarily well approximated by

$$\frac{d}{d\alpha} E_x^\alpha f(x^\Delta(t)) \Big|_{\alpha=\alpha_0}$$

for large t and small Δ . It is such approximations that are actually used in the applications. It is important to note that for large enough t , the quality of the approximation is uniformly good in (small) Δ .

THEOREM 7.4. *Suppose that Assumptions 2.1–2.5 and 6.1 hold. Then for $f \in L^\infty(R^r)$,*

$$\begin{aligned} (7.10) \quad \langle \mu_\alpha(\alpha_0), f \rangle &= \lim_{\Delta \rightarrow 0} \langle \mu_\alpha^\Delta(\alpha_0), f \rangle \\ &= \lim_{\substack{\Delta \rightarrow 0 \\ t \rightarrow \infty}} \int \mu^\Delta(dx, \alpha_0) \frac{d}{d\alpha} E_x^\alpha f(x^\Delta(t)) \Big|_{\alpha=\alpha_0}, \end{aligned}$$

where the limits as $\Delta \rightarrow 0$, $t \rightarrow \infty$ can be taken in any way at all.

Proof. Write, by the invariance of $\mu^\Delta(\alpha)$ and the differentiability

$$\begin{aligned} (7.11) \quad \frac{d}{d\alpha} \langle \mu^\Delta(\alpha), f \rangle_{\alpha=\alpha_0} &= \langle \mu_\alpha^\Delta(\alpha_0), P^\Delta(\alpha_0, t)f \rangle \\ &\quad + \langle \mu^\Delta(\alpha_0), P_\alpha^\Delta(\alpha_0, t)f \rangle, \end{aligned}$$

where $P^\Delta(\alpha_0, t)f(x) = E_x^{\alpha_0} f(x^\Delta(t))$ and $t > 0$. We have $P^\Delta(\alpha_0, t)f \rightarrow \tilde{f}^{\alpha_0} = \langle \mu(\alpha_0), f \rangle$ as $\Delta \rightarrow 0$, $t \rightarrow \infty$. Also, $\{\mu_\alpha^\Delta(\alpha_0), \Delta > 0\}$ is of bounded variation by the corollary to Theorem 7.2. Thus $\langle \mu_\alpha^\Delta(\alpha_0), P^\Delta(\alpha_0, t)f \rangle \rightarrow 0$ as $\Delta \rightarrow 0$ and $t \rightarrow \infty$, which yields the theorem. \square

A pathwise result. With the approximation of Theorem 7.4, we can give the pathwise result. Since we only have one long realization and cannot explicitly calculate the derivatives of the expectations, we need to show that a long simulation of $\{X_n^\Delta, n < \infty\}$ can yield a good approximation to the right side of (7.10) for fixed Δ . Typically, the t_0 in Theorem 7.5 is as large as can be, consistent with a modest sample variance.

THEOREM 7.5. *Suppose that Assumptions 2.1–2.5 and 6.1 hold. Fix $t = n\Delta$. Let $f(\cdot)$ be bounded and continuous. Then as $T \rightarrow \infty$ (or with centered f used as discussed in §5)*

$$(7.12) \quad \begin{aligned} & \frac{1}{T} \int_0^T [Z^\Delta(t_0 + s, \alpha_0) - Z^\Delta(s, \alpha_0)] f(x^\Delta(t_0 + s)) ds \\ & \xrightarrow{P} \int \mu^\Delta(dx, \alpha_0) E_{x^\Delta}^{\alpha_0} Z^\Delta(t_0, \alpha_0) f(x^\Delta(t_0)) \\ & = \left\langle \mu^\Delta(\alpha_0), \frac{d}{d\alpha} E^\alpha f(x^\Delta(t_0)) \Big|_{\alpha=\alpha_0} \right\rangle. \end{aligned}$$

Proof. Fix t_0 . Define $\delta Z^\Delta(t_0, s) = Z^\Delta(t_0 + s, \alpha_0) - Z^\Delta(s, \alpha_0)$ and $Y^\Delta(t_0, s) = \delta Z^\Delta(t_0, s) f(x^\Delta(t_0 + s))$. Then the process (parameter T) defined by

$$M^\Delta(T) = \int_0^T [Y^\Delta(t_0, s) - E_{x^\Delta(s)}^{\alpha_0} Y^\Delta(t_0, s)] ds$$

is a zero mean martingale whose variance is $O(T)$. Thus Kronecker's lemma implies that $M^\Delta(T)/T \rightarrow 0$ with probability one. This implies that for the purpose of evaluating the limit of the left side of (7.12), we can replace it by

$$(7.13) \quad \frac{1}{T} \int_0^T q(x^\Delta(s)) ds,$$

where we define

$$q(x^\Delta(s)) = E_{x^\Delta(s)}^{\alpha_0} Y^\Delta(t_0, s) = \frac{d}{d\alpha} E_{x^\Delta(s)}^\alpha f(x^\Delta(t_0 + s)) \Big|_{\alpha=\alpha_0}.$$

The function $q(\cdot)$ is continuous and bounded. Then, the ergodic properties of $\{X_n^\Delta, n < \infty\}$ imply that (as $T \rightarrow \infty$) (7.13) converges with probability one to its mean value $\langle \mu^\Delta(\alpha_0), q \rangle$, which is just the center term of (7.12). \square

8. Numerical comparisons. The approximation method of §7 has been simulated and compared with alternative methods on a variety of problems of dimension up to seven. Here, we comment on some comparisons with a finite difference method. The alternative methods are all described and discussed in [1], and we will repeat only a few of the comments made there.

The basic method used for all methods takes one long simulation, over an interval T_1 . A basic estimation interval T_0 is given, and the approximate model $X^\Delta(\cdot)$ is simulated. $N = T_1/T_0$ estimates of the derivative are made in the long simulation interval, each using T_0 units of time. Let X_n^Δ denote the state of the system at the start of the n th subinterval. Then X_n^Δ is the initial condition for the estimate on the $(n+1)$ st subinterval. The detailed results reported here are for a two-dimensional problem, with the parameter α being a scalar. We comment on larger problems later. For the finite difference estimate, a pair of simulations must be taken, with a parameter set at $\alpha_0 \pm \delta\alpha$, for some small $\delta\alpha$. The samples of the δw in (5.1) for the second member of the pair was the same as that of the first member of the pair, with the samples being independent from pair to pair. This reduced the variance over what would have been the case if all the samples of the δw random variables had been mutually independent, as in [1]. The reduction was particularly large if the system

TABLE 8.1

$T_0 = 3$		
Finite Difference ($\delta\alpha = .05$)		
	sample mean	sample standard deviation
derivative	.168	.247
cost	.363	.149
CPU Time		32.04
 AC		
	sample mean	sample standard deviation
derivative	.164	.216
cost	.364	.127
CPU Time		18.9
 Weighted AC (Derivative only)		
	sample mean	sample standard deviation
λ		
.1	.160	.19
.5	.153	.14
CPU Time		20.1

was linear, and the cost function smooth, although there was a noticeable reduction in the variance in all cases tested.

The two-dimensional problem was the noise-driven Van der Pol equation

$$dx_1 = x_2 dt, \quad dx_2 = [10x_2(1 - x_1^2) - \alpha x_1]dt + dw,$$

where $\alpha_0 = 2$. Note that this system is degenerate. Nevertheless, the method works well. The cost function of interest was $\int_0^S k(x(s))ds/S$ for large S , where

$$k(x) = I_{\{|x_2| \geq 0.3\}}.$$

The simplest estimator is

$$(8.1) \quad \frac{1}{N} \sum_{n=1}^N \frac{1}{T_0} \int_{nT_0}^{nT_0+T_0} [Z^\Delta(s, \alpha_0) - Z^\Delta(nT_0, \alpha_0)] k(X^\Delta(s)) ds.$$

An “antithetic” variable method was always used, since it gives a reduced variance: Let N be an even number, and let the δw samples used for the $2n$ th estimate be the negative of that used for the $(2n-1)$ th ($n = 1, 2, \dots$) estimate, with the δw used for the $(2n-1)$ th estimates ($n = 1, 2, \dots, N/2$) being mutually independent.

The centered form, where $k(X^\Delta(s))$ is replaced by the centered $k(X^\Delta(s)) - \bar{k}(nT_0 + T_0)$, where the centering is a sample estimate of the value of the cost at the cited time, actually gave better results. This method is referred to as the AC-method in the tables below (antithetic variable, centered). The centering is zero mean, but helps reduce the variance. As $n \rightarrow \infty$, (8.1) converges to

$$\frac{d}{d\alpha} \int \mu(dx, \alpha_0) E_x^{\alpha_0} k(X^\Delta(T_0)).$$

TABLE 8.2

$T_0 = 10$		
Finite Difference ($\delta\alpha = .05$)		
	sample mean	sample standard deviation
derivative	.157	.243
cost	.364	.052
CPU Time		104.8
AC		
	sample mean	sample standard deviation
derivative	.162	.304
cost	.364	.032
CPU Time		65.5
Weighted AC (Derivative only)		
	sample mean	sample standard deviation
$\lambda = .5$.157	.106
$\lambda = 1$.150	.07
CPU Time		68.3

For large enough T_0 , this is a good estimate of the desired derivative. A better procedure would be to divide the interval $[0, T_0]$ into a reasonable number of subintervals to get a better approximation to the first centered form discussed in §5. We must keep in mind, however, that the CPU time required for a large number of subdivisions might be better used for taking more samples.

A third method, called the weighted AC-method, often (but not always) was advantageous. As $s \rightarrow \infty$, the variance of $[Z^\Delta(s, \alpha_0) - Z^\Delta(nT_0, \alpha_0)]$ goes to ∞ . If the system has a “short” memory, then the “earlier” part of the $Z^\Delta(\cdot)$ process contributes little to the estimate in the following sense: Let $nT_0 + T_0 > s > s_0 > nT_0$, and write

$$\begin{aligned}
 & [Z^\Delta(s, \alpha_0) - Z^\Delta(nT_0, \alpha_0)]k(X^\Delta(s)) \\
 &= [Z^\Delta(s, \alpha_0) - Z^\Delta(s_0, \alpha_0)]k(X^\Delta(s)) \\
 & \quad + [Z^\Delta(s_0, \alpha_0) - Z^\Delta(nT_0, \alpha_0)]k(X^\Delta(s)).
 \end{aligned}$$

Then the mean value of the second term goes to zero as $s - s_0 \rightarrow \infty$. If we reduce the sample interval, however, then a bias is added. To balance the opposing effects, we use a weighted substitute \tilde{Z}^Δ for Z^Δ , constructed as follows, where $\lambda \in (0, 1)$ is a weighing factor or exponential discount of the past (notation for the nondegenerate case):

$$\tilde{Z}^\Delta((n+1)\Delta) = [\sigma^{-1}(X_i^\Delta, \alpha_0)b_\alpha(X_i^\Delta, \alpha_0)]' \delta w(i\Delta) - \lambda \tilde{Z}^\Delta(n\Delta).$$

For the problem reported on here, this method gave excellent results. In other cases, where the “approach to ergodicity” is slower, a substantial bias could be introduced into the estimates.

Refer to Tables 8.1–8.3, where the sample means of the derivative estimates, their sample standard deviations, and the required CPU time are given. For the finite

TABLE 8.3

$$T_0 = 20$$

Finite Difference ($\delta\alpha = .05$)

	sample mean	sample standard deviation
derivative	.168	.246
cost	.365	.032
CPU Time		209.8

AC

	sample mean	sample standard deviation
derivative	.168	.537
cost	.365	.021
CPU Time		65.5

Weighted AC (Derivative only)

	sample mean	sample standard deviation
$\lambda = .5$.154	.058
$\lambda = 1$.163	.101
CPU Time		137.05

difference estimates, $N=2500$ was used, and $N=5000$ otherwise. This is because two system simulations per finite difference estimate are needed, and only one for our method. The important quantity, however, is the sample standard deviation per CPU time unit. Note that the sample standard deviation for the weighted AC-method decreases as T_0 increases, while that for the AC-method increases. We can readily see the advantages of the methods introduced here. For linear systems, the finite difference method seems to work better owing to the 'smoothness' of the dependence of the estimates on the noise, and the value of the difference interval was not too important (did not seriously affect the sample variance), as long as it was small enough to control the bias.

There are important dimensionality advantages to our methods. Suppose that the dimension of the parameter is m . Then, to get a single estimate of a gradient, a finite difference method needs to simulate the system either $(m + 1)$ or $2m$ times, depending on the finite difference method used (one-sided or central). Our method requires the simulation of only one sample path per estimate, and the calculation of one Z -variable per component of the parameter. The calculation of the Z -variable, however, is usually much simpler than doing a simulation of the system. This is particularly true if the system is of high dimension, or if the dynamical terms are hard to compute. Thus, our methods require much less computer time than does the finite difference method, particularly for high-dimensional and nonlinear problems. Alternative methods, such as the finite difference method, can compensate for this only by having a better quality estimate, i.e., one with smaller bias or sample variance.

We emphasize that no general rule has been found that can tell us which method would be preferable for any particular class of problems. For many problems, the alternative methods are preferable. All methods must be taken as serious candidates, and techniques sought for their realization so that they perform as well as possible.

REFERENCES

- [1] JICHUAN YANG AND HAROLD J. KUSHNER, *A Monte Carlo method for the sensitivity analysis and parametric optimization of nonlinear stochastic systems*, LCDS report 89-90, Brown University, Providence, RI, SIAM J. Control Optim., 29 (1991), pp. 1216–1249.
- [2] R. Z. KHAZMINSKII, *Ergodic properties of recurrent diffusions and stabilization of the solution to the cauchy problem for parabolic systems*, Theory Probab. Appl., 5 (1960), pp. 179–196. (English translation.)
- [3] H. J. KUSHNER, *Optimality conditions for the average cost per unit time problem with a diffusion model*, SIAM J. Control Optimization, 16 (1978), pp. 330–346.
- [4] S. OREY, *Limit Theorems for Markov Chain Transition Probabilities*, Van Nostrand, London, 1971.
- [5] H. J. KUSHNER, *Approximation and Weak Convergence Methods for Random Processes, with Applications to Stochastic Systems Theory*, 1971. M.I.T. Press, Cambridge, MA, 1984.
- [6] S. N. ETHIER AND T. G. KURTZ, *Markov Processes: Characterization and Convergence*, John Wiley, New York, 1986.
- [7] H. J. KUSHNER, *Converse theorem for stochastic Lyapunov functions*, SIAM J. Control Optim., 5 (1967), pp. 228–233.

CONVEX DUALITY AND GENERALIZED SOLUTIONS IN OPTIMAL CONTROL PROBLEM FOR STOPPED PROCESSES: THE DETERMINISTIC MODEL*

HANG ZHU†

Abstract. The approach consists of imbedding the original control problem tightly in a convex mathematical programming problem on the space of measures and then solving the latter problem by duality in convex analysis. The dual to the control problem is to find the supremum of all smooth subsolutions to the Bellman equation. Because of the effect of stops at the boundary of the domain, a different formulation of strong and weak problems will be adopted to make use of the convex duality method. The results about the decomposition of weak measures provide a clear interpretation for such a effect in the weak formulation of our control problem.

Key words. optimal control, stopped processes, smooth subsolution, convex duality, strong and weak problems

AMS(MOS) subject classifications. 49J27, 49N15

1. Introduction. Recently, a new method has been developed in the study of optimal control problems. The method, which originated in Young's concept of generalized curves, took the approach from a point of view of convex analysis, and, in particular, use of duality between spaces of continuous functions and spaces of measures. This so-called "convex duality" method was first introduced by Vinter and Lewis [9], [10] in their study, which concerns deterministic control problems. They established the equivalence between the control problem and a related "weak" problem. This weak problem is actually a convex mathematical programming problem, and the objects in the weak problem are called the "generalized solutions." Upon establishing the equivalence of two problems, the new optimality conditions could then be obtained by setting up the Fenchel dual of the weak problem. It turns out that the dual problem is actually to find the supremum among all smooth subsolutions of the Bellman equation. Later on, this method was greatly simplified by Fleming and Vermes and applied in their studies of optimal control problems of (piecewise) deterministic processes [2], [8] as well as diffusion processes [3], [4]. The results in those papers were proved for the fixed finite time horizon and the infinite time horizon discounted cost problems.

In this paper, we consider a deterministic optimal control problem for stopped processes and address the role that the boundary will play in the duality formulation of Fleming–Vermes approach. By the "stopped process," we mean that the state process X_s under the control will be stopped at the first time it hits the boundary of a given domain $\Omega \subset \mathbb{R}^N$. Our objective is to minimize the cost functional that has a nonnegative integral cost up to the stopping time. Because of the effect of stops at the boundary of the domain, the correct weak problems are defined by coupling both the "interior" measure and "boundary" measure together in our formulation so that the standard Fenchel duality of convex analysis (see the Appendix in the end of paper) can still be set up appropriately to assert that the dual of our control problem is again to find the supremum among all smooth subsolutions of the Bellman equation.

*Received by the editors January 24, 1990; accepted for publication (in revised form) February 21, 1991.

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

The equivalence between strong and weak formulations is obtained in §3 where we have presented a simple new proof for our main equivalence theorem. The proof is based on two important results, called the “canonical extension” and the “canonical extraction,” respectively. These two results give us an intuitive portrait about the structural content of weak measures in terms of the “stopping” effect in our control problems.

Related problems could be found in Soner [7] and Vermes [8]. In [7], where the author studies the control problem with state constraints, the processes are pushed back continuously from the boundary of the domain, and the value function there is characterized in terms of the notion of constrained viscosity solution. In [8], the processes can jump back from the boundary into the interior of the domain, and the author proved a disintegration theorem for general occupation measures and obtained, by the same convex duality method, a characterization of the value function as the supremum of all smooth subsolutions to the Bellman equation. The assumptions in both papers assure that all processes have a uniformly nonvanishing lateral speed in the boundary region, which is basically equivalent to our assumption (H3). The present paper differs from [7], [8] in the sense that the processes under control will simply stop when they come to the boundary region. The major achievement of this paper is that it establishes a two-way link between weak measures generated by processes stopped at the boundary and measures generated by continuing processes in the closed domain.

2. The stopped processes problem and its strong/weak formulations.

Consider the following deterministic control system,

$$(2.1) \quad \begin{aligned} \dot{X}_s &= f(s, X_s, u_s) \\ X_t &= x, \quad 0 \leq t \leq s \leq T, \end{aligned}$$

where the control u_s is Borel measurable, taking values in the control space U , and f is an R^N -valued function defined on $[0, T] \times R^N \times U$.

We want to restrict the state process of (2.1) to stay inside some given bounded domain Ω in R^N , and will stop the process X_s whenever it first hits the boundary $\partial\Omega$. The objective is to minimize

$$(2.2) \quad \mathcal{J}^u(t, x) = \int_t^\tau l_0(s, X_s, u_s) ds + L_0(\tau, X_\tau),$$

where the running cost l_0 and the terminal cost L_0 are the functions defined on $[t, T] \times R^N \times U$ and $[t, T] \times R^N$, respectively, and τ is the stopping time defined as

$$\tau = \begin{cases} \inf\{s \geq t : X_s \notin \Omega\} \\ T, & \text{if } X_s \in \Omega, \quad \forall t \leq s \leq T. \end{cases}$$

Note that both X_s and τ depend on the initial condition (t, x) and the control u_s being applied. In the rest of our paper, we will work with this cost functional (2.2) except in Theorem 3.4 (the main equivalence result), where $l_0 \geq 0$ and $L_0 \equiv 0$ are assumed.

Then the value function is given by

$$(2.3) \quad \psi(t, x) = \inf_{u \in \mathcal{U}} \mathcal{J}^u(t, x),$$

for $(t, x) \in \Sigma = [0, T] \times \bar{\Omega}$. Here $\mathcal{U} = \mathcal{U}(t)$ denotes the class of all Borel measurable, U -valued functions u_s on $[t, T]$. The Bellman equation takes the form

$$(2.4) \quad \inf_{u \in \mathcal{U}} \{ \partial_t \psi + f \cdot \nabla_x \psi + l_0 \} = 0, \quad \text{in } [t, T] \times \Omega \times U$$

with the terminal-boundary condition $\psi(t, x) = L_0(t, x)$ on $\Sigma_\partial = [0, T] \times \partial\Omega \cup \{T\} \times \bar{\Omega}$. This equation is understood in the viscosity solution sense of Lions [5].

It is convenient to summarize here the following notation and assumptions:

$$\Sigma = [0, T] \times \bar{\Omega}, \quad S = \Sigma \times U, \quad \Sigma_\partial = [0, T] \times \partial\Omega \cup \{T\} \times \bar{\Omega}.$$

$C(\Sigma), C(S), C(\Sigma_\partial)$ = the spaces of real-valued continuous functions on $\Sigma, S, \Sigma_\partial$, respectively.

$$C^{1,1}(\Sigma) = \{ \phi \in C(\Sigma) : \phi_t, \phi_{x_i} \in C(\Sigma), 1 \leq i \leq N \}.$$

(H1) The bounded domain $\Omega \subset \mathbb{R}^N$ has a C^1 -smooth boundary $\partial\Omega$. The control space U is a compact subset in \mathbb{R}^M .

(H2) The running cost function $l_0 \in C(S)$, $l_0 \geq 0$ and the terminal cost function $L_0 \in C(\Sigma_\partial)$. The system function $f \in C(S)$ satisfies the uniform x -Lipschitz condition

$$|f(t, x_1, u) - f(t, x_2, u)| \leq K|x_1 - x_2|, \quad \forall x_1, x_2 \in \bar{\Omega},$$

where K is some constant independent of (t, u) .

(H3) The system (2.1) assumes the following “boundary controllability” condition:

$$(2.5) \quad 0 \in \text{int}\{f(t, x, U)\}, \quad \forall (t, x) \in [0, T] \times \partial\Omega.$$

Here $\text{int}\{f(t, x, U)\}$ stands for the interior of the set $f(t, x, U) = \{f(t, x, u) : u \in U\}$.

Let us now start our formulation of strong and weak problems.

We introduce a more elaborate state space \tilde{S} , which is composed of the “interior” space S and the “boundary” space Σ_∂ as two separate components. Then every function $\tilde{l} \in C(\tilde{S})$ is denoted as a pair (l, L) with $l \in C(S)$ and $L \in C(\Sigma_\partial)$. The dual space $\mathcal{M}(\tilde{S})$ will consist of all pairs of measures $\tilde{M} = (M, N)$ with $M \in \mathcal{M}(S)$ and $N \in \mathcal{M}(\Sigma_\partial)$, where $\mathcal{M}(S), \mathcal{M}(\Sigma_\partial)$ are spaces of finite, signed measures on S, Σ_∂ , respectively.

Given the initial condition (t, x) , each control $u \in \mathcal{U}$ defines a measure $\tilde{M}^u \in \mathcal{M}(\tilde{S})$ as follows:

$$(2.6) \quad \begin{aligned} \langle \tilde{l}, \tilde{M}^u \rangle &= \langle l, M^u \rangle + \langle L, N^u \rangle \\ &= \int_t^\tau l(s, X_s^u, u_s) ds + L(\tau, X_\tau), \quad \forall \tilde{l} = (l, L) \in C(\tilde{S}). \end{aligned}$$

The measure $\tilde{M}^u = (M^u, N^u)$ clearly has the following disintegration:

$$\begin{aligned} M^u(ds, dx, du) &= I_{[t, \tau)} \lambda(ds) \times \delta_{X_s}(dx) \times \delta_{u_s}(du) \\ N^u(ds, dx) &= \delta_{(\tau, X_\tau)}(ds, dx), \end{aligned}$$

where X_s^u is the state process of (2.1) when the control u is applied, τ is the stopping time of this X_s^u . Here λ is the Lebesgue measure on $[0, T]$, $\delta_{(\tau, X)}$ is the Dirac measure at (τ, X) , and $I_{[t, \tau)}$ is an indicator function. The measures M^u and N^u are called,

respectively, the “interior” measure and the “boundary” measure generated by the underlying process.

By the above identification of u with $\tilde{M}^u = (M^u, N^u)$, our control problem can then be reformulated as a linear minimization problem:

$$(2.7) \quad \text{Minimize} \quad \langle \tilde{l}_0, \tilde{M} \rangle = \int_{\tilde{S}} \tilde{l}_0 d\tilde{M}$$

over $\mathcal{M}^s(t, x) = \{\tilde{M}^u = (M^u, N^u) : u \in \mathcal{U}\}$. This is called the *strong problem*.

For each $\tilde{M}^u = (M^u, N^u)$, we can calculate the norms

$$\begin{aligned} \|M^u\| &= \langle 1, M^u \rangle = \tau - t \leq T - t \\ \langle L, N^u \rangle &= L(\tau, X_\tau) \leq \|L\| \Rightarrow \|N^u\| \leq 1, \quad \forall u \in \mathcal{U}. \end{aligned}$$

It follows from the fundamental theorem of calculus that for any $\phi \in C^{1,1}(\Sigma)$

$$(2.8) \quad \phi(\tau, X_\tau) - \phi(t, x) = \int_t^\tau A\phi(s, X_s, u_s) ds, \quad \forall u \in \mathcal{U},$$

where $A\phi(t, x, u) = \partial_t \phi(t, x) + f(t, x, u) \cdot \nabla_x \phi(t, x)$. If we define the operator

$$\tilde{A} : C^{1,1}(\Sigma) \rightarrow C(\tilde{S})$$

by

$$\tilde{A}\phi(\tilde{x}) = \begin{cases} A\phi(t, x, u), & \text{if } \tilde{x} = (t, x, u) \in S \\ -\phi(t, x), & \text{if } \tilde{x} = (t, x) \in \Sigma_\partial, \end{cases}$$

then the above formula can be rewritten in terms of (2.6)

$$\begin{aligned} \phi(t, x) &= -\langle A\phi, M^u \rangle + \langle \phi, N^u \rangle \\ &= -\langle \tilde{A}\phi, \tilde{M}^u \rangle, \quad \forall \phi \in C^{1,1} \text{ and } u \in \mathcal{U}. \end{aligned}$$

Given the initial data (t, x) , let us introduce the space

$$(2.9) \quad \mathcal{M}^w(t, x) = \left(\tilde{M} = (M, N) \in \mathcal{M}(\tilde{S}) : \begin{array}{l} 1) M, N \geq 0 \text{ and } \|M\| \leq T - t, \|N\| \leq 1 \\ 2) \phi(t, x) = -\langle \tilde{A}\phi, \tilde{M} \rangle, \forall \phi \in C^{1,1}(\Sigma) \end{array} \right).$$

Note that this set is both convex and w^* -compact in $\mathcal{M}(\tilde{S})$.

The *weak problem* is therefore formulated as: Given the initial condition (t, x) and $\tilde{l}_0 \in C(\tilde{S})$,

$$(2.10) \quad \text{Minimize} \quad \langle \tilde{l}_0, \tilde{M} \rangle = \int_{\tilde{S}} \tilde{l}_0 d\tilde{M}$$

over $\mathcal{M}^w(t, x)$.

Throughout this paper, we use the following notation:

$$\psi^s(t, x) = \inf_{\mathcal{M}^s(t, x)} \langle \tilde{l}_0, \tilde{M} \rangle, \quad \psi^w(t, x) = \inf_{\mathcal{M}^w(t, x)} \langle \tilde{l}_0, \tilde{M} \rangle.$$

The weak formulation is reduced to the Fenchel's problem in convex analysis. By solving its dual, a maximization problem in the space $C(\tilde{S})$, it turns out that the

dual of minimization (2.10) is to find the supremum of all smooth subsolutions to the Bellman equation (2.4). This is the standard duality result in [4] and its modification will be demonstrated in the Appendix.

3. Decomposition of the weak measures and the main equivalence theorem. We establish two results in this section. The first is the canonical extension theorem, which constructs the weak measure of a continuing process from a stopped one. The second result is that from every weak measure, a minimal, stopped part can be extracted. As a consequence of these results, we give a short proof for the main equivalence theorem.

To present our results, we need to consider the corresponding deterministic control problem with state constraints. Let \bar{S} be the state space, which consists of the “interior” space S and the “terminal” space $\Sigma_T = \{T\} \times \bar{\Omega}$, as two separate components. The formulation of strong and weak problems for the state-constraint problem can then be defined in a similar way to those in §2.

We denote the strong space

$$(3.1) \quad \mathcal{M}^s(t, x) = \{(M_c^u, N_c^u) : X_s^u \in \bar{\Omega}, \forall s \in [t, T]\},$$

where each $\bar{M}_c^u = (M_c^u, N_c^u)$ is associated to a control u , which keeps the state process X_s^u inside $\bar{\Omega}$ for the entire period $[t, T]$, by the formula

$$\langle \bar{l}, \bar{M}_c^u \rangle = \int_t^T l(s, X_s^u, u_s) ds + L(T, X_T), \quad \forall \bar{l} = (l, L) \in C(\bar{S}).$$

If the process X_s^u hits the boundary $\partial\Omega$ at first time $\tau < T$, we have a stopped process that corresponds to a measure $\tilde{M}^u = (M^u, N^u) \in \mathcal{M}^s(t, x)$, and it is associated to the same control u by the (2.6)

$$\langle \tilde{l}, \tilde{M}^u \rangle = \int_t^\tau l(s, X_s^u, u_s) ds + L(\tau, X_\tau), \quad \forall \tilde{l} = (l, L) \in C(\tilde{S}).$$

Note that N_c^u , as a measure on Σ_T , can be identified with a measure on Σ_∂ by a trivial extension since $\Sigma_T \subset \Sigma_\partial$. By such an identification, we have that $\bar{M}_c^u \in \mathcal{M}^w(t, x)$ but not necessarily in $\mathcal{M}^s(t, x)$.

Compared to the original formulation in §2, we have the fundamental theorem of calculus:

$$\phi(T, X_T) - \phi(t, x) = \int_t^T A\phi(s, X_s, u_s) ds, \quad \forall \phi \in C^{1,1}(\Sigma).$$

By introducing the operator $\bar{A} : C^{1,1}(\Sigma) \rightarrow C(\bar{S})$ as

$$\bar{A}\phi(\bar{x}) = \begin{cases} A\phi(t, x, u), & \text{if } \bar{x} = (t, x, u) \in S \\ -\phi(t, x), & \text{if } \bar{x} = (t, x) \in \Sigma_T \end{cases}$$

we then define the weak space for the state-constraint problem:

$$(3.2) \quad \mathcal{M}_c^w(t, x) = \left(\begin{array}{l} \bar{M}_c = (M_c, N_c) \in \mathcal{M}(\bar{S}) : \\ 1) M_c, N_c \geq 0 \text{ and } \|M_c\| = T - t, \|N_c\| = 1 \\ 2) \phi(t, x) = -\langle \bar{A}\phi, \bar{M}_c \rangle, \forall \phi \in C^{1,1}(\Sigma) \end{array} \right).$$

Clearly, $\mathcal{M}_c^w(t, x) \subset \mathcal{M}^w(t, x)$ in the above-mentioned sense. This is expected since the “terminal” space Σ_T is contained in the boundary space $\Sigma_\partial = [0, T] \times \partial\Omega \cup \Sigma_T$.

Let us now illustrate how the measures in $\mathcal{M}^s(t, x)$ are related to the measures in $\mathcal{M}_c^s(t, x)$. For each measure $\tilde{M} = (M, N) \in \mathcal{M}^s(t, x)$, there exists a control u_s such that with $L \equiv 0$, (2.6) becomes

$$\langle l, M \rangle = \int_t^\tau l(s, X_s^u, u_s) ds, \quad \forall l \in C(S),$$

where X_s^u is the state process under the control u_s and τ is the stopping time of this X_s^u .

If $\tau < T$, the measure M corresponds to a stopped process X_s^u . In this case, a “canonical” extension M_c of M can be defined as

$$(3.3) \quad \langle l, M_c \rangle = \langle l, M \rangle + \int_\tau^T l(s, \xi, \bar{u}_s) ds,$$

where the control \bar{u}_s is chosen by the following lemma to keep the stopped process X_s^u at the place $\xi = X_\tau^u \in \partial\Omega$ until the terminal time T . That is, $X_s^{\bar{u}} = \xi$ for $\tau \leq s \leq T$.

LEMMA 3.1. *For each $(\tau, \xi) \in [t, T] \times \partial\Omega$, there is a control \bar{u}_s such that the corresponding state process $\bar{X}_s = X_s^{\bar{u}}$ satisfies*

$$\bar{X}_s = \xi, \quad \forall \tau \leq s \leq T.$$

Furthermore, this \bar{u}_s can be chosen to depend Borel-measurably on (τ, ξ) .

Proof. We consider the constant process $\bar{X}_s = X_s(\tau, \xi) \equiv \xi$, for $\tau \leq s \leq T$. By assumption (H3), this \bar{X}_s satisfies

$$\dot{X}_s(\tau, \xi) \equiv 0 \in f(s, X_s(\tau, \xi), U), \quad \forall \tau \leq s \leq T.$$

Now invoking the Filippov lemma (see the generalized version in McShane and Warfield [6]), a Borel-measurable function $\bar{u}_s = u_s(\tau, \xi) \in U$ can be chosen such that it also depends on (τ, ξ) Borel-measurably, and it holds

$$\dot{X}_s(\tau, \xi) \equiv 0 = f(s, X_s(\tau, \xi), u_s), \quad \forall \tau \leq s \leq T.$$

Therefore, we have $\bar{X}_s = X_s^{\bar{u}}$ and certainly $\bar{X}_s = \xi$ for $\tau \leq s \leq T$. \square

Motivated by our illustration before Lemma 3.1, let us prove the first result, which constructs a “canonical” extension for each weak measure.

PROPOSITION 3.2 (Canonical extension). *Every $\tilde{M} = (M, N) \in \mathcal{M}^w(t, x)$ has an extension $\bar{M}_c = (M_c, N_c) \in \mathcal{M}_c^w(t, x)$.*

Proof. For each $(\tau, \xi) \in [t, T] \times \partial\Omega$, we define $(M_{\tau, \xi}, N_{\tau, \xi}) \in \mathcal{M}_c^s(t, x)$ as follows:

$$\begin{aligned} \langle l, M_{\tau, \xi} \rangle &= \int_\tau^T l(s, \xi, \bar{u}_s) ds, \quad \forall l \in C([0, T] \times \partial\Omega \times U) \\ \langle L, N_{\tau, \xi} \rangle &= L(T, \xi), \quad \forall L \in C(\Sigma_T). \end{aligned}$$

By Lemma 3.1, the control \bar{u}_s is chosen to depend Borel-measurably on (τ, ξ) , then $M_{\tau, \xi}, N_{\tau, \xi}$ depends on (τ, ξ) Borel-measurably, too. We can therefore define the measures

$$(3.4) \quad \begin{aligned} M_\partial &= \int_{[t, T] \times \partial\Omega} M_{\tau, \xi} N(d\tau, d\xi) \\ N_\partial &= \int_{[t, T] \times \partial\Omega} N_{\tau, \xi} N(d\tau, d\xi) \end{aligned}$$

and take

$$M_c = M + M_\partial, \quad N_c = N|_{\Sigma_T} + N_\partial.$$

Note that each $M_{\tau,\xi}$ acts only on the boundary space $[t, T] \times \partial\Omega \times U$, then so does M_∂ .

Clearly, $M_c, N_c \geq 0$ and $M_c \in \mathcal{M}(S), N_c \in \mathcal{M}(\Sigma_T)$.

Let us compute that for any $\phi \in C^{1,1}$,

$$\begin{aligned} \langle A\phi, M_\partial \rangle &= \int_{[t,T] \times \partial\Omega} \langle A\phi, M_{\tau,\xi} \rangle N(d\tau, d\xi) \\ &= \int_{[t,T] \times \partial\Omega} \langle \phi, N_{\tau,\xi} \rangle N(d\tau, d\xi) - \langle \phi, N \rangle + \int_{\Sigma_T} \phi dN \\ &= \langle \phi, N_c \rangle - \langle \phi, N \rangle \end{aligned}$$

and then

$$\begin{aligned} \phi(t, x) &= -\langle A\phi, M \rangle + \langle \phi, N \rangle \\ &= -\langle A\phi, M_c \rangle + \langle A\phi, M_\partial \rangle + \langle \phi, N \rangle \\ &= -\langle A\phi, M_c \rangle + \langle \phi, N_c \rangle = -\langle \bar{A}\phi, \bar{M}_c \rangle \end{aligned}$$

Next by setting $\phi \equiv 1$ in the formula $\phi(t, x) = -\langle \bar{A}\phi, \bar{M}_c \rangle$, we get

$$1 = \langle 1, N_c \rangle \Rightarrow \|N_c\| = 1,$$

and by setting $\phi(t, x) = T - t$ in the same formula and noting that $\langle \phi, N_c \rangle = T - T = 0$, we have

$$T - t = \langle 1, M_c \rangle \Rightarrow \|M_c\| = T - t.$$

Hence $\bar{M}_c = (M_c, N_c) \in \mathcal{M}_c^w(t, x)$, and it is called in our terminology the “canonical” extension of $\bar{M} = (M, N)$. \square

Let “cl” denote the closure in the w^* -topology of $\mathcal{M}(\bar{S})$, consider also the space of relaxed measures in the state-constraint problem

$$\mathcal{M}_c^g(t, x) = cl\mathcal{M}_c^s(t, x),$$

and denote $\mathcal{M}_0^g(t, x) = \{M_c : (M_c, N_c) \in \mathcal{M}_c^g(t, x)\}$. Clearly, this is a w^* -closed set in $\mathcal{M}(\bar{S})$. We observe that for each $\mu \in \mathcal{M}_0^g(t, x)$, there exists a relaxed control ν_s such that $\mu = M^\nu$ has the disintegration

$$\mu(dr, dx, du) = \lambda(dr) \times \delta_{X_r^\nu}(dx) \times \nu_r(du).$$

The notion of relaxed control was introduced in Fleming [1], it is a Borel-measurable, $\mathcal{P}(U)$ -valued process. Here $\mathcal{P}(U)$ = the space of all probability measures on U . Correspondingly the state process X_s^ν is required to satisfy, instead of (2.1), that

$$(3.5) \quad \dot{X}_s = \int_U f(s, X_s, u) \nu_s(du).$$

In the following arguments, we will associate each μ to a state process X_s^μ , which is the solution of the above relaxed system under the relaxed control ν_s . In terms of $\mu \in \mathcal{M}_0^g(t, x)$, we can rewrite (3.5) as follows:

$$X_s^\mu = x + \int_t^s \int_U \int_{\bar{\Omega}} f(r, x, u) \mu(dr, dx, du), \quad \text{for } t \leq s \leq T.$$

Then it is clear that X_s^μ depends continuously on μ with respect to the w^* -topology of $\mathcal{M}_0^g(t, x)$.

We now prove the second result, which extracts a minimal, stopped part from each weak measure.

PROPOSITION 3.3 (Canonical extraction). *For each $\tilde{M} = (M, N) \in \mathcal{M}^w(t, x)$, there exists a measure $\tilde{M}^0 = (M^0, N^0) \in \mathcal{M}^w(t, x)$ such that $M^0([t, T] \times \partial\Omega \times U) = 0$ and $M^0 \leq M$.*

Remark. This M^0 is called the “stopped” part of M . It is generated by the processes stopped at the boundary of the domain.

Proof. We have shown in Proposition 3.2 that $\tilde{M} = (M, N) \in \mathcal{M}^w(t, x)$ has a “canonical” extension $\tilde{M}_c = (M_c, N_c) \in \mathcal{M}_c^w(t, x)$. By the formula below (3.8) in Fleming [2], this M_c has an integral representation:

$$M_c = \int_{\mathcal{M}_0^g(t, x)} \mu \alpha(d\mu)$$

for some probability measure α on $\mathcal{M}_0^g(t, x)$.

For each $\mu = M^\nu \in \mathcal{M}_0^g(t, x)$, we define the measure μ^0 by

$$\langle l, \mu^0 \rangle = \int_t^\tau \int_U l(s, X_s^\mu, u) \nu_s(du) ds, \quad \forall l \in C(S),$$

where τ is the stopping time of X_s^μ . In short, $d\mu^0 = I_{[t, \tau)} d\mu$.

Since X^μ depends continuously in μ , the stopping time $\tau^\mu = \inf\{s \geq t : X_s^\mu \notin \Omega\}$ is then lower semicontinuous with respect to μ in the w^* -topology. Let $\mu_n = M^{\nu_n} \in \mathcal{M}_0^g(t, x)$ such that $\mu_n \xrightarrow{w^*} \mu$, and $\mu^0(\mu_n)$ satisfies

$$\begin{aligned} \langle l, \mu^0(\mu_n) \rangle &= \int_t^{\tau^n} \int_U l(s, X_s^n, u) \nu_s^n(du) ds \\ (3.6) \quad &= \int_t^{\tau^n} \int_U \int_\Omega l(s, x, u) \mu^n(ds, dx, du), \quad \forall l \in C(S), \end{aligned}$$

where X_s^n is the state process associated to μ^n and τ^n is the stopping time of X_s^n . By the fact $\tau \leq \liminf_{n \rightarrow \infty} \tau^n$, we pass the limit in (3.6) to get

$$\langle l, \mu^0(\mu) \rangle \leq \liminf_{n \rightarrow \infty} \langle l, \mu^0(\mu_n) \rangle, \quad \text{for } l \geq 0.$$

However, any $l \in C(S)$ can be written as $l = l^+ - l^-$, then $\langle l, \mu^0(\cdot) \rangle = \langle l^+, \mu^0(\cdot) \rangle - \langle l^-, \mu^0(\cdot) \rangle$ is the difference of two lower semicontinuous functions and therefore Borel-measurable as a function of μ , for any $l \in C(S)$. Equivalently, $\mu^0(\cdot)$ is Borel-measurable in μ with respect to the w^* -topology of $\mathcal{M}_0^g(t, x)$.

We are then allowed to define

$$(3.7) \quad M^0 = \int_{\mathcal{M}_0^g(t, x)} \mu^0 \alpha(d\mu).$$

Because each μ^0 has the properties that $\mu^0 \leq \mu$ and $\mu^0([t, T] \times \partial\Omega \times U) = 0$, hence $M^0 \leq M_c$ and $M^0([t, T] \times \partial\Omega \times U) = 0$, too. That is, M^0 vanishes on the boundary space $[t, T] \times \partial\Omega \times U$.

Next, we compute that for any $\phi \in C^{1,1}$,

$$\begin{aligned}
 \langle A\phi, M^0 \rangle &= \int_{\mathcal{M}_0^g(t,x)} [\phi(\tau, X_\tau^\mu) - \phi(t, x)] \alpha(d\mu) \\
 (3.8) \quad &= \int_{\mathcal{M}_0^g(t,x)} \phi(\tau, X_\tau^\mu) \alpha(d\mu) - \phi(t, x) = \langle \phi, N^0 \rangle - \phi(t, x),
 \end{aligned}$$

where the boundary measure N^0 is defined by

$$(3.9) \quad \langle L, N^0 \rangle = \int_{\mathcal{M}_0^g(t,x)} L(\tau, X_\tau^\mu) \alpha(d\mu), \quad \forall L \in C(\Sigma_\partial).$$

This is well defined since both τ and X_τ^μ depend on $\mu \in \mathcal{M}_0^g(t, x)$ Borel-measurably.

Clearly, $M^0, N^0 \geq 0$, and $\|M^0\| \leq T - t, \|N^0\| \leq 1$. Together with (3.8), it implies that $\tilde{M}^0 = (M^0, N^0) \in \mathcal{M}^w(t, x)$ by its definition.

Finally, to show that $M^0 \leq M$, we note that $M^0 \leq M_c$ and $M|_{[t,T) \times \Omega \times U} = M_c|_{[t,T) \times \Omega \times U}$, since $M_c = M + M_\partial$, and M_∂ acts only on the boundary space Σ_∂ . Then, we have

$$\begin{aligned}
 \langle l, M^0 \rangle &= \int_{[t,T) \times \Omega \times U} l dM^0 \leq \int_{[t,T) \times \Omega \times U} l dM_c \\
 &= \int_{[t,T) \times \Omega \times U} l dM \leq \langle l, M \rangle
 \end{aligned}$$

for any $l \geq 0$ in $C(S)$. \square

The next result is entitled the Equivalence Principle in the convex duality approach for the optimal control problems. Owing to the results established above, we are able to give a simple proof here without directly employing the convolution argument in Fleming and Vermes [2], [3], [4].

THEOREM 3.4. *Assume that $l_0 \geq 0, L_0 \equiv 0$, then $\psi^w(t, x) = \psi^s(t, x)$. That is, the strong formulation (2.7) is equivalent to the weak formulation (2.10) in our stopped processes problem. Moreover, the minimum ψ is attained by a “stopped” measure.*

As a consequence of this equivalence theorem and Theorem A.3 in the Appendix, we conclude that

$$\psi(t, x) = \sup\{\phi(t, x) : \phi \text{ is a smooth subsolution}\}.$$

Proof. First, $\psi^w(t, x) \leq \psi^s(t, x)$ holds trivially since $\mathcal{M}^w(t, x) \supset \mathcal{M}^s(t, x)$.

To prove the opposite inequality, let us take any $\tilde{M} = (M, N) \in \mathcal{M}^w(t, x)$. By Proposition 3.3, there is a $\tilde{M}^0 = (M^0, N^0) \in \mathcal{M}^w(t, x)$ such that M^0 is a minimal, stopped part of M in terms of the boundary $\partial\Omega$.

Since $l_0 \geq 0$ and $L_0 \equiv 0$, it follows from the integral representation (3.7) that

$$\begin{aligned}
 \langle \tilde{l}_0, \tilde{M} \rangle &= \langle l_0, M \rangle \geq \langle l_0, M^0 \rangle \\
 &= \int_{\mathcal{M}_0^g(t,x)} \langle l_0, \mu^0 \rangle \alpha(d\mu) \\
 &\geq \int_{\mathcal{M}_0^g(t,x)} \psi^s \alpha(d\mu) = \psi^s(t, x).
 \end{aligned}$$

Therefore, we have $\psi^w(t, x) \geq \psi^s(t, x)$. \square

The proof also implies that the minimum ψ must be attained by some “stopped” measure. Otherwise, by employing the argument in Proposition 3.3, the “stopped” part can always be extracted so that it will yield a smaller cost. \square

Remark. (1) It is indicated in [2] that $\mathcal{M}_c^w(t, x)$ = the w^* -convex closure of $\mathcal{M}_c^s(t, x)$. We conjecture that in our stopped processes problem

$$(3.10) \quad \mathcal{M}^w(t, x) = \text{the } w^* - \text{convex closure of } \mathcal{M}^s(t, x)$$

holds true, although we have not obtained a rigorous proof for this assertion.

(2) The weak space $\mathcal{M}^w(t, x)$ in our stopped processes problem has measures generated by the stopped processes as well as measures generated by the continuing processes. Compared to the weak space $\mathcal{M}_c^w(t, x)$ in the state-constraint problem, this is a larger space and therefore yields a smaller minimum (i.e., the value function), which is expected in our problem since we have a nonnegative penalty cost.

Appendix. Convex duality and smooth subsolutions of the Bellman equation. The results in this section, which exhibits the approach from a viewpoint of convex analysis, are very similar to §5 of Vermes [8] and §4 of Fleming and Vermes [4]. Most arguments are the duplicates of those in [4], [8], therefore we will only state the results and indicate the differences in our proofs.

Let us bring the weak problem (2.10) into the Fenchel normal form.

Define the function

$$\tilde{h}(\tilde{M}) = \begin{cases} \langle \tilde{l}_0, \tilde{M} \rangle, & \tilde{M} \in \mathcal{M}^w(t, x) \\ +\infty, & \tilde{M} \in \mathcal{M}(\tilde{S}) \setminus \mathcal{M}^w(t, x) \end{cases}$$

for $\tilde{l}_0 = (l_0, L_0) \in C(\tilde{S})$. Clearly, this extended real-valued function \tilde{h} is convex and lower semicontinuous, and its dual \tilde{h}^* is defined by the Legendre–Fenchel transformation

$$\tilde{h}^*(\tilde{l}) = \sup_{\tilde{M} \in \mathcal{M}(\tilde{S})} \{ \langle \tilde{l}, \tilde{M} \rangle - \tilde{h}(\tilde{M}) \}, \quad \forall \tilde{l} \in C(\tilde{S}).$$

Rather than attempting to find \tilde{h}^* , we write $\tilde{h} = \tilde{h}_1 - \tilde{h}_2$ in such a way that the duals $\tilde{h}_1^*, \tilde{h}_2^*$ can be found explicitly.

Introduce the following pair of spaces:

$$\mathcal{M}_1(t, x) = \{ \tilde{M} = (M, N) \in \mathcal{M}(\tilde{S}) : 1) \text{ holds in (2.9)} \}$$

$$\mathcal{M}_2(t, x) = \{ \tilde{M} = (M, N) \in \mathcal{M}(\tilde{S}) : 2) \text{ holds in (2.9)} \}$$

and define

$$\tilde{h}_1(\tilde{M}) = \begin{cases} \langle \tilde{l}_0, \tilde{M} \rangle, & \tilde{M} \in \mathcal{M}_1(t, x) \\ +\infty, & \tilde{M} \in \mathcal{M}(\tilde{S}) \setminus \mathcal{M}_1(t, x), \end{cases}$$

$$\tilde{h}_2(\tilde{M}) = \begin{cases} 0, & \tilde{M} \in \mathcal{M}_2(t, x) \\ -\infty, & \tilde{M} \in \mathcal{M}(\tilde{S}) \setminus \mathcal{M}_2(t, x). \end{cases}$$

Both \tilde{h}_1 and $-\tilde{h}_2$ are clearly convex and lower semicontinuous, since $\mathcal{M}_1(t, x)$ and $\mathcal{M}_2(t, x)$ are both convex and closed. Then, it is immediate that the weak problem is equivalent to the Fenchel problem:

Minimize $\tilde{h}_1(\tilde{M}) - \tilde{h}_2(\tilde{M})$ over $\tilde{M} \in \mathcal{M}(\tilde{S})$. In other words,

$$\min_{\mathcal{M}^w(t, x)} \{ \langle \tilde{l}_0, \tilde{M} \rangle \} = \min_{\mathcal{M}(\tilde{S})} \{ \tilde{h}_1(\tilde{M}) - \tilde{h}_2(\tilde{M}) \}.$$

The Legendre–Fenchel transforms of \tilde{h}_1, \tilde{h}_2 are computed in the following lemma.
LEMMA A.1.

$$\begin{aligned} (1) \quad \tilde{h}_1^*(\tilde{l}) &= (T-t)\|(l-l_0)^+\| + \|(L-L_0)^+\|, \quad \forall \tilde{l} = (l, L) \in C(\tilde{S}) \\ (2) \quad \tilde{h}_2^*(\tilde{l}) &= \begin{cases} -\lim \phi_n(t, x), & \text{if } \tilde{l} = \lim_{n \nearrow \infty} \tilde{A}\phi_n \text{ with } \phi_n \in C^{1,1}(\Sigma) \\ -\infty, & \text{otherwise} \end{cases} \end{aligned}$$

The proof of this lemma will follow exactly the same line as that in Fleming and Vermes [4], [8].

We recall the Bellman equation (2.4):

$$\inf_{u \in U} \{ \partial_t \psi + f \cdot \nabla_x \psi + l_0 \} = 0$$

with the terminal and boundary conditions $\psi(t, x) = L_0(t, x)$ on Σ_∂ .

DEFINITION A.2. The function ϕ is called a smooth subsolution to the Bellman's equation (2.4) if $\phi \in C^{1,1}(\Sigma)$ satisfies

$$A^u \phi + l_0 \geq 0, \quad \text{in } [0, T) \times \Omega \times U$$

and $\phi(t, x) \leq L_0(t, x)$ on Σ_∂ .

The next theorem is the main result of this convex duality approach. Roughly it states that seeking maximal solution of the Bellman equation is the dual to the weak problem formulated in the §2. Under the current weak assumptions, no classical solutions to the Bellman equation need to exist.

THEOREM A.3. For any $\tilde{l}_0 = (l_0, L_0)$ in $C(\tilde{S})$, we have

$$\begin{aligned} \psi^w(t, x) &\stackrel{\text{def}}{=} \min \{ \langle \tilde{l}_0, \tilde{M} \rangle : \tilde{M} \in \mathcal{M}^w(t, x) \} \\ &= \sup \{ \phi(t, x) : \phi \text{ is a smooth subsolution} \}. \end{aligned}$$

That is, the value function (i.e., the minimum) of the weak problem is the upper envelope (i.e., supremum) of the smooth subsolutions to the Bellman equation.

Proof. As in Fleming and Vermes [4], [8], substitute \tilde{h}_1^* and \tilde{h}_2^* of Lemma A.1 in the Rockafellar duality formula and use the fact that $\{\tilde{A}\phi : \phi \in C^{1,1}\}$ is dense in set $\{\tilde{l} : \tilde{h}_2^*(\tilde{l}) > -\infty\}$. We obtain

$$\begin{aligned} \psi^w(t, x) &= \min \{ \tilde{h}_1(\tilde{M}) - \tilde{h}_2(\tilde{M}) : \tilde{M} \in \mathcal{M}(\tilde{S}) \} \\ &= \sup \{ \tilde{h}_2^*(\tilde{l}) - \tilde{h}_1^*(\tilde{l}) : \tilde{l} \in C(\tilde{S}) \} \\ &= \sup \{ \phi(t, x) - (T-t)\|(A\phi + l_0)^-\| - \|(\phi - L_0)^+\| : \phi \in C^{1,1} \}. \end{aligned}$$

Let $\Phi(t, x) = \phi(t, x) - (T-t)\|(A\phi + l_0)^-\| - \|(\phi - L_0)^+\|$. Clearly $\Phi \in C^{1,1}(\Sigma)$ and it satisfies

$$\begin{aligned} A\Phi + l_0 &= (A\phi + l_0) + \|(A\phi + l_0)^-\| \geq 0, \quad \text{in } [0, T) \times \Omega \times U, \\ \Phi(t, x) &\leq \phi - \|(\phi - L_0)^+\| \leq L_0(t, x), \quad \text{on } \Sigma_\partial. \end{aligned}$$

By Definition A.2, Φ is indeed a smooth subsolution to the Bellman equation (2.4) and we conclude the result of theorem. \square

Acknowledgment. The author would like to give his sincere thanks to Professor Wendell H. Fleming for suggesting the problem, his good advice, and his careful

reading of this manuscript, which brought a substantial change of the early version. The author also would like to thank the referees for many very helpful comments.

REFERENCES

- [1] W. H. FLEMING, *Generalized solutions in optimal stochastic control*, Presented at the second Kingston Conference on Differential Games and Control Theory, June 1976.
- [2] ———, *Generalized solutions and convex duality in optimal control*, in PDE and Calculus of Variation, Vol.1, F. Colombini, et al, eds., 1989, pp. 461–471.
- [3] W. H. FLEMING AND D. VERMES, *Generalized solutions in the optimal control of diffusions*, in IMA Volumes in Math and Application, W. H. Fleming and P. L. Lions, eds., Springer-Verlag, Berlin, New York, 1987, pp. 119–127.
- [4] ———, *Convex duality approach to the optimal control of diffusions*, SIAM J. Control, 27 (1989), pp. 1136–1155.
- [5] P. L. LIONS, *Generalized Solutions of Hamilton–Jacobi Equations*, Pitman Research Notes, London, (1982).
- [6] E. J. MCSHANE AND R. B. WARFIELD, *On a Filippov's implicit functions lemma*, Proc. Amer. Math. Soc., 18 (1967), pp. 41–47.
- [7] H. M. SONER, *Optimal control with state-space constraint I*, SIAM J. Control, 24 (1986), pp. 552–566.
- [8] D. VERMES, *Optimal control of piecewise deterministic Markov process*, Stochastics, 14 (1985), pp. 546–570.
- [9] R. B. VINTER AND R. M. LEWIS, *The equivalence of strong and weak formulations for certain problems in optimal control*, SIAM J. Control, 16 (1978), pp. 546–570.
- [10] ———, *A necessary and sufficient condition for optimality of dynamic programming type, making no a priori assumptions on the controls*, SIAM J. Control, 16(1978), pp. 571–583.

MESH INDEPENDENCE OF THE GRADIENT PROJECTION METHOD FOR OPTIMAL CONTROL PROBLEMS*

C. T. KELLEY[†] AND E. W. SACHS[‡]

Abstract. The gradient projection method and its application to optimal control have been analyzed with regard to the rate of convergence quite extensively. Here another aspect of this method is considered. In finite dimension the set of active indices is identified after finitely many iterations under mild nondegeneracy assumptions. However, it is clear that this property is restricted to the finite-dimensional case. In this paper, for example, a sequence of discretized optimal control problems is considered, and it is observed that the number of steps to identify all active indices increases with the refinement of the discretization. A result analogous to the finite-dimensional result is valid in this situation if identification of active indices is understood in the correct light. This paper shows that if a different termination criterion is imposed, then the number of necessary steps for termination is indeed mesh-independent. Numerical observations illustrating this result are reported for various examples from optimal control.

Key words. gradient projection method, finite identification of active indices, mesh independence, optimal control problems

AMS(MOS) subject classifications. 49D05, 49D07, 65K10

1. Introduction. Consider the constrained optimization problem,

$$(1.1) \quad \min F(u) \text{ such that } u \in U,$$

where H is a Hilbert space, $U \subset H$ is a bounded, closed, and convex set, and $F : H \rightarrow \mathbb{R}$ is the objective function. A popular method for solving (1.1) numerically is the gradient projection method. This algorithm requires an inexpensive way to compute the projection onto U , $P : H \rightarrow U$, where

$$(1.2) \quad P(h) \text{ solves } \min_{u \in U} \|u - h\|.$$

The sequence of iterates is defined by

$$(1.3) \quad u_{k+1} = P(u_k - \alpha_k \nabla F(u_k)),$$

where $\alpha_k > 0$ is a stepsize parameter that is determined by some stepsize rule. We will consider the Armijo stepsize rule, shown below.

For fixed $\tilde{\alpha} > 0, \beta, \delta \in (0, 1)$ set $\alpha_k = \tilde{\alpha}$ if

$$F(u_k) - F(P(u_k - \tilde{\alpha} \nabla F(u_k))) \geq \delta \langle \nabla F(u_k), u_k - P(u_k - \tilde{\alpha} \nabla F(u_k)) \rangle.$$

Otherwise, choose α_k such that $\alpha_k = \beta^{m_k} \tilde{\alpha}$, where m_k is the smallest integer $m \in \mathbb{N}$ with

$$(1.4) \quad F(u_k) - F(P(u_k - \beta^m \tilde{\alpha} \nabla F(u_k))) \geq \delta \langle \nabla F(u_k), u_k - P(u_k - \beta^m \tilde{\alpha} \nabla F(u_k)) \rangle.$$

* Received by the editors April 16, 1990; accepted for publication (in revised form) March 29, 1991.

[†] North Carolina State University, Department of Mathematics, Box 8205, Raleigh, North Carolina 27695-8205. The research of this author was supported by National Science Foundation grants DMS-8900410, INT-8800560, and Air Force Office of Scientific Research grant AFOSR-ISSA-890044.

[‡] Universität Trier, FB IV–Mathematik, Postfach 3825, 5500 Trier, Germany.

In Hilbert space this algorithm has been considered in [7], [10], and [14] with application to control problems. In recent years the gradient projection method has again become the focus of various researchers for problems with simple constraints, called bound constraints or box constraints. In the finite-dimensional case, $H = \mathbb{R}^D$; this means that

$$(1.5) \quad U = \{u \in \mathbb{R}^D : a_i^- \leq u_i \leq a_i^+, i = 1, \dots, D\},$$

with fixed vectors $a^-, a^+ \in \mathbb{R}^D$ such that $a_i^- < a_i^+, i = 1, \dots, D$. This type of structure occurs, in particular, for the discrete formulations of optimal control problems with constraints on the magnitude of controls: Let $x(t) \in \mathbb{R}$ and $u(t) \in \mathbb{R}$ denote the state and control, respectively. The state $x(t)$ depends on the control through the differential equation

$$(1.6) \quad \dot{x}(t) = f(t, x(t), u(t)), \quad x(0) = x_0 \in \mathbb{R},$$

where $f \in C^1(\mathbb{R}^3)$. The feasible set of controls U is defined as

$$U = \{u \in L_\infty[0, T] : a^-(t) \leq u(t) \leq a^+(t) \text{ a.e. on } [0, T]\},$$

where $a^-, a^+ \in L_\infty[0, T]$ are fixed functions such that $a^-(t) < a^+(t)$ almost everywhere $[0, T]$.

The function that is to be minimized is given by

$$(1.7) \quad \int_0^T L(t, x(t), u(t)) dt,$$

with $L \in C^1(\mathbb{R}^3)$.

In discretized form, the optimal control problem can be written as

$$(1.8) \quad \min \sum_{i=1}^D L_i(x_i, u_i),$$

where x depends on u through a discretization of (1.6) such as

$$(1.9) \quad x_{i+1} = x_i + f_i(x_i, u_i), \quad i = 0, 1, \dots, D$$

and

$$(1.10) \quad L_i, f_i \in C^1(\mathbb{R}^2), \quad i = 0, 1, \dots, D$$

are functions arising from the discretization process. Our analysis does not require one-step schemes such as (1.9), and the discretization of the differential equation could also be realized through a finite element approach.

The results in this paper will require that the solutions to the discrete problems be in the same space as those of the continuous problem. This can be done for the discrete optimal control problems discussed above by regarding the control, state, and adjoint vectors as step functions. Full details of this correspondence are given in §4.

An index of a constraint is considered active when the inequality is an equality. Bertsekas [2] pointed out that an advantage of the gradient projection method is that it allows us to add or to drop several active indices in one iteration. This is especially important for problems with many active constraints at the solution. Furthermore,

Bertsekas showed that under certain nondegeneracy and second-order sufficiency assumptions, the set of active indices is identified after finitely many iterations.

Nondegeneracy assumption. Let the gradients of the constraints corresponding to the active indices at a point u_* , satisfying the necessary optimality conditions, be linearly independent and let the Lagrange multipliers corresponding to active indices be positive.

THEOREM 1.1 (Bertsekas [2]). *If u_k is a sequence produced by the gradient projection method with the Armijo step-size rule that converges to a solution u_* , where the nondegeneracy assumption holds, then there exists an index k_0 such that the set of active indices of u_k remains unchanged for k larger than k_0 .*

In a series of papers, the convergence analysis for the gradient projection method has been investigated. Dunn [7] related the convergence rate to a condition on the growth of the objective function or its first-order approximation as we move away from the optimal point. For some $a > 0$ and $\nu \geq 1$, let

$$(1.11) \quad \gamma(\sigma) \geq a\sigma^\nu \quad \text{for } \sigma > 0$$

hold where

$$(1.12) \quad \gamma(\sigma) := \inf_{u \in U, \|u - u_*\| \geq \sigma} F(u) - F(u_*), \quad \sigma > 0.$$

Note that the growth of γ could come from F or the feasible set U . A special case of Theorem 4.3 in [7] is the following result.

THEOREM 1.2. *Let U be a closed convex bounded subset of H and let F be convex and bounded with a continuous Fréchet derivative. If with γ in (1.12) $\gamma(\sigma) > 0$ for $\sigma > 0$, then for $r_k = F(u_k) - F(u_*)$ we have $r_k = o(1/k)$. In particular,*

$$r_k \leq r_0 \left(1 + \frac{4r_0\delta\alpha}{D^2}k \right)^{-1}$$

for $k \in \mathbb{N}$ with $\alpha_k \geq \alpha > 0$ and $D = \max\{\|u - v\| : u, v \in U\}$. If for some $a > 0$ (1.11) holds, then for $\nu = 2$ there exists $\hat{\alpha} \in (0, 1)$ with

$$\frac{r_{k+1}}{r_k} \leq \hat{\alpha} \text{ for } k \in \mathbb{N}$$

and, for $\nu > 2$,

$$r_k = O\left(\frac{1}{k^{\nu(\nu-2)}}\right).$$

The boundedness away from 0 for the stepsizes is a common requirement in this context, and we refer to Lemma 3.4. Papers [2], [3], [8], and [9] considered Newton-type variations of the gradient projection method to improve the convergence behavior. In [4] the notion of the projected gradient was used to improve and extend the result on finite identification of active constraints. In all papers, the choice of the stepsize rule plays a quite important role. References [5], [15], and [16] deal with a trust-region approach for the gradient projection method as a globalization strategy. Also in these publications, the finite identification of active constraints is a major issue in the proofs.

In the numerical results of [2] it is mentioned that for discretized optimal control problems the number of iterates required to identify the set of active constraints

TABLE 1.1
Finite identification of active indices

D_N	k_{act} for Ex. 1	k_{act} for Ex. 2	k_{act} for Ex. 3
4	163	56	6
8	779	53	7
16	5032	76	14
32	36472	127	24
64	277582	231	56

increases as the discretization is refined. This can lead to a prohibitively large number of iterates until the minimum is achieved. We illustrate this for some examples, which we use later in the section on numerical results (§4). We choose

$$L(t, x, u) = \frac{1}{2}x^2, \quad f(t, x, u) = bu, \quad x_0 = 1, \quad T = 1, \quad a^+(\cdot) = -a^-(\cdot) = 0.5$$

and three choices for b :

Example 1.1. $b(t) = (t - 0.5)^3$;

Example 1.2. $b(t) = (t - 0.5)^3 + 0.03$;

Example 1.3. $b(t) = \cos(3t)$.

We discretize this optimal control problem with (1.8) and (1.9):

$$D = D_N, \quad L_i(x, u) = \frac{1}{2D_N}x^2, \quad f_i(x, u) = b(i/D_N)u.$$

The set of active indices is defined as

$$I(u) = \{i : u_i = a_i^- \text{ or } u_i = a_i^+\}.$$

It can be shown that all the active indices are identified after finitely many steps. However, as we can see from Table 1.1, this statement can become irrelevant in the sense that the number of active indices and the number of steps required for their identification increases without bound as the discretization is refined. The number of steps needed to identify $D_N - 1$ active indices

$$k_{\text{act}} = \min\{k \in \mathbb{N} : |I(u_k)| = D_N - 1\}$$

is shown in Table 1.1.

The first example does not satisfy the quadratic growth condition, (1.11) with $\nu = 2$, but the second, which is a slightly perturbed example does satisfy the quadratic growth condition, albeit with a small σ . Table 1.1 clearly shows that the influence of an infinite-dimensional ill posed problem affects the algorithmic behavior of the finite-dimensional discretized problem. As a well-posed example, we use Example 1.3, where (1.11) is satisfied.

All three examples in Table 1.1 show that the number of steps required to identify the active indices is dependent on the dimension of the problem. It is evident (see, e.g., [16]) that for optimization problems with infinitely many constraints the active set will not be identified in finitely many steps. For discretizations of such problems, more iterations are necessary to identify the active set as the discretization is refined. This is an unsatisfactory situation because the computational effort does not only become larger because of the increasing dimension but also because of the higher number of iterations required for active constraint identification. Moreover, in the context of control, much of this work resolves the solution of the discrete problem to an accuracy beyond that of the discretization of the differential equation, and is

wasted. In this paper we will give measures of constraint identification that limit the number of iterations in a way that is independent of the dimension.

We seek mesh-independence results of several kinds for the identification of the active set of indices. One of the concepts of mesh independence that we use here was first defined in [1] and [12] in the context of nonlinear equations. In those papers, iteration schemes for solution of $F(u) = 0$ in a Hilbert space H were compared with those for an approximate equation $F_N(u) = 0$ on a space H_N . The number of iterates required to drive the norm of F or F_N to a given size were compared. Define $F_\infty = F$, $H_\infty = H$, and

$$l_N(\varepsilon) = \min\{k : \|F_N(u_k^N)\|_{H_N} < \varepsilon\}, \quad l_\infty(\varepsilon) = \min\{k : \|F(u_k)\|_H < \varepsilon\},$$

where u_k^N and u_k are the k th iterate for the iteration scheme on H_N and H , respectively. The results in [1] and [12] required only mild assumptions on the methods and the types of approximate problems and stated that for every $\varepsilon > \delta > 0$ sufficiently small there was $N_{\varepsilon,\delta}$ such that if $N > N_{\varepsilon,\delta}$, then

$$l_\infty(\varepsilon + \delta) \leq l_N(\varepsilon) \leq l_\infty(\varepsilon).$$

Our goal in this paper is to formulate and prove results of this type in the context of approximations to (1.1) in the specific case of optimal control with constraints on the magnitude of controls.

Section 2 contains the definition of the set J_ε , which is a subset of the set of active indices. This set is identified after finitely many steps, even for a problem in function space. For example, if the gradient satisfies a condition that plays an important role in the convergence rate analysis, then the set J_ε converges also at a certain rate to the set of active indices. Section 3 is devoted to approximate problems and the statements and proofs of the mesh-independence results. Under certain conditions on the convergence of the gradients, we can prove that the number of steps necessary to identify the set J_ε is independent of the meshsize, as is the number of iterates needed to drive the norm of the projected gradient to below ε . These statements illustrate the type of mesh independence that we seek.

An important assumption is that the stepsizes are uniformly bounded away from zero. We verify this assumption for an Armijo-type stepsize rule. In §4 we consider two examples from optimal control. For a problem with ordinary differential equations, we give numerical results that justify the claim that the number of iterations needed to identify all active constraints can be excessively high. The results indicate that the number of iterations remains almost constant under a mesh refinement if other termination criteria are used. The second example from the optimal control of the heat equation through boundary input gives similar results. It is noted that these observations are important in particular when the optimization problems are illposed.

2. Identification of active indices. In this section we consider statements about the finite identification of active indices for the infinite-dimensional problem (1.2) in function space. For simplicity, we choose in the definition of the feasible set U the functions a^- , a^+ to be 1 and -1:

$$(2.1) \quad U = \{u \in L_\infty[0, T] : |u(t)| \leq 1 \text{ a.e. on } [0, T]\}.$$

Since the projection in a Hilbert space is essential in the framework of the algorithm, the analysis for the convergence is usually carried with respect to the L_2 norm. The Fréchet differentiability in this norm has to be checked for the specific examples

because it is a much stronger requirement than differentiability with respect to the L_∞ norm. We denote the gradient $\nabla F(u_k)$ by s_k . In the context of the optimal control problems (1.6), (1.7) the gradient can be computed, if it exists, by

$$s_k(t) = L_u(t, x_k(t), u_k(t)) + p_k(t)f_u(t, x_k(t), u_k(t)),$$

where p_k solves

$$-\dot{p}_k = L_x(x_k, u_k) + p_k f_x(x_k, u_k), \quad p_k(T) = 0.$$

To formulate a result on finite identification of active indices in function space, we reconsider the bang-bang principle. Since at the optimal control u_*

$$\int_0^T s_*(t)(u(t) - u_*(t)) dt \geq 0 \quad \text{for all } u \in U,$$

where s_* is defined analogously to s_k , we deduce that

$$u_*(t) = -\operatorname{sgn} s_*(t) \quad \text{for a.a. } t \in [0, T], \quad \text{where } |s_*(t)| > 0.$$

Obviously, in a numerical scheme, the positivity of $s_*(t)$ will be considered satisfied if for an appropriate small value $\varepsilon > 0$, the inequality

$$(2.2) \quad |s_*(t)| \geq \varepsilon$$

holds. We will show in the following lemma that for all t where (2.2) is true, the values of u_* are identified after a fixed iteration index. This allows us to prove a result on finite identification of indices t , which satisfy (2.2), and to refine this statement later to mesh-independence results for the discretized problems.

LEMMA 2.1. *Let $s_k, s_*, u_k, u_* \in L_\infty[0, T]$, and $\alpha_k \in \mathbb{R}$ be given with some $\alpha > 0$ such that*

$$(2.3) \quad \alpha_k \geq \alpha > 0 \quad \text{for all } k \in \mathbb{N}$$

and

$$(2.4) \quad u_{k+1}(t) = \operatorname{sat}(u_k(t) - \alpha_k s_k(t)) \quad \text{a.e. on } [0, T].$$

In (2.4) the function sat is defined in the usual way as

$$(2.5) \quad \operatorname{sat}(u) = \begin{cases} 1 & \text{if } u > 1, \\ u & \text{if } u \in [-1, 1], \\ -1 & \text{if } u < -1. \end{cases}$$

Suppose that

$$\lim_{k \rightarrow \infty} \|s_k - s_*\|_{L_\infty[0, T]} = 0,$$

and, for all $\varepsilon > 0$,

$$(2.6) \quad |u_*(t)| = 1 \quad \text{a.e. on } J_\varepsilon = \{t \in [0, T] : |s_*(t)| \geq \varepsilon\}.$$

Then, for all $\varepsilon > 0$, there exists $k_\varepsilon \in \mathbb{N}$ with

$$(2.7) \quad u_k(t) = u_*(t) \quad \text{for all } k \geq k_\varepsilon, \quad t \in J_\varepsilon.$$

Proof. Choose $k_1(\varepsilon) \in \mathbb{N}$ such that, for all $k \geq k_1(\varepsilon)$,

$$\|s_k - s_*\|_{L_\infty[0,T]} \leq \frac{\varepsilon}{2}.$$

We introduce the sets

$$(2.8) \quad J_\varepsilon^+ = \{t \in J_\varepsilon : s_*(t) \geq \varepsilon\} \quad \text{and} \quad J_\varepsilon^- = \{t \in J_\varepsilon : s_*(t) \leq -\varepsilon\}.$$

Then

$$(2.9) \quad s_k(t) \geq \frac{\varepsilon}{2} \quad \text{for all } k \geq k_1(\varepsilon) \text{ and a.e. on } J_\varepsilon^+.$$

Hence for $l(\varepsilon) \in \mathbb{N}$ with $l(\varepsilon) \geq 4/(\alpha\varepsilon)$ we have with (2.3)

$$(2.10) \quad \sum_{k=k_1(\varepsilon)}^{l(\varepsilon)+k_1(\varepsilon)} \alpha_k s_k(t) \geq l(\varepsilon) \alpha \frac{\varepsilon}{2} \geq 2 \quad \text{a.e. on } J_\varepsilon^+,$$

and hence, according to iteration rule (2.4) of the gradient projection method,

$$(2.11) \quad u_k(t) = -1 \quad \text{for } k \geq l(\varepsilon) + k_1(\varepsilon) + 1 \text{ a.e. on } J_\varepsilon^+$$

Similarly, we obtain

$$(2.12) \quad u_k(t) = +1 \quad \text{for } k \geq l(\varepsilon) + k_1(\varepsilon) + 1 \text{ a.e. on } J_\varepsilon^-;$$

i.e., (2.7) holds with $k_\varepsilon = k_1(\varepsilon) + l(\varepsilon) + 1$. \square

It is evident that the size of the interval J_ε where finite identification occurs depends on the shape of the graph of s_* . If s_* has only finitely many zeros in $[0, T]$, then the length $m(J_\varepsilon)$ of J_ε converges to T , $\lim_{\varepsilon \rightarrow 0} m(J_\varepsilon) = T$, but this convergence can be arbitrarily slow. The following lemma gives conditions that relate the rate at which $m(J_\varepsilon)$ increases to properties of s^* . We define the set of zeros of s_* by $\mathcal{Z} = \{t \mid s_*(t) = 0\}$ and the complement by

$$(2.13) \quad \mathcal{A} = \{t \mid s_*(t) \neq 0\}.$$

We use the following condition

$$(2.14) \quad \begin{aligned} &\mathcal{Z} \text{ is a finite union of intervals } [c_i, d_i], i = 1, \dots, N_Z, \\ &\text{and there is } \hat{\sigma} > 0 \text{ and } \mu \geq 1 \text{ such that} \end{aligned}$$

$$|s_*(t)| \geq \hat{\sigma} \text{dist}(t, \mathcal{Z})^\mu.$$

LEMMA 2.2. *If (2.14) holds, then*

$$(2.15) \quad m(\mathcal{A} - J_\varepsilon) \leq c \varepsilon^{1/\mu}$$

holds for all $\varepsilon > 0$ sufficiently small and some fixed $c > 0$.

Proof. If $t \in \mathcal{A} - J_\varepsilon$, then

$$\varepsilon > |s_*(t)| > \hat{\sigma} \text{dist}(t, \mathcal{Z})^\mu.$$

Hence, if

$$\overline{\mathcal{Z}} = \bigcup_{1 \leq i \leq N_Z} [c_i, d_i],$$

then, if $\delta = (\varepsilon/\hat{\sigma})^{1/\mu}$,

$$t \in \bigcup_{1 \leq i \leq N_Z} ((c_i - \delta, c_i] \cup [d_i, d_i + \delta)).$$

This completes the proof with $c = 2N_Z\hat{\sigma}^{-1/\mu}$. \square

3. Approximate problems and mesh independence. In this section we approximate the problem of minimizing F on a subset U of H by the minimization of some $F_N : H_N \rightarrow \mathbb{R}$ on a set $U_N \subset H_N$. The issue in this section is how measures of progress toward identification of the set of active constraints \mathcal{A} vary with the level of approximation. Such questions come under the category of mesh-independence results.

We consider three notions of mesh independence. The first is the assertion in Theorem 3.1. This says that the iterate k after which the active interval has been identified up to ε is independent of N . This can be seen in the numerical results in Table 4.2.

We show that the set J_ε of indices is identified after a number of steps that is independent of the discretization level N in the following sense. In this theorem it is assumed that s_k^N converges uniformly in L_∞ to s_* , which is checked in §4 for the examples under consideration. This assumption also holds if the uniform rate estimate on the function values in Theorem 1.2 is applied to the following estimates. Conditions on f and L that yield the convergence of u_*^N, x_*^N, s_*^N to u_*, x_*, s_* like linearity of f and convexity of L , (cf. [6, p. 116]) imply that, for some $c > 0$,

$$c\|u_k^N - u_*^N\|_{L_2} \leq F_N(u_k^N) - F_N(u_*^N).$$

Hence u_k^N converges uniformly to u_* and, therefore, also to the corresponding switching functions.

THEOREM 3.1. *Let $s_k^N, s_* \in L_\infty[0, T]$, $k, N \in \mathbb{N}$ be such that*

$$(3.1) \quad \lim_{N, k \rightarrow \infty} \|s_k^N - s_*\|_{L_\infty[0, T]} = 0$$

holds and that there is α such that

$$(3.2) \quad \alpha_k^N \geq \alpha > 0.$$

Then for all $\varepsilon > 0$ such that (2.6) is true, there exist $k_\varepsilon, N_\varepsilon \in \mathbb{N}$ with

$$(3.3) \quad u_k^N(t) = u_*(t) \quad \text{for all } k \geq k_\varepsilon, N \geq N_\varepsilon, \text{ and } t \in J_\varepsilon.$$

Proof. Choose $k_1(\varepsilon), N_\varepsilon \in \mathbb{N}$ such that for all $k \geq k_1(\varepsilon)$, $N \geq N_\varepsilon$,

$$\|s_k^N - s_*\|_{L_\infty[0, T]} \leq \varepsilon/2.$$

Define $J_\varepsilon^+, J_\varepsilon^-$ as in (2.8). If we consider s_k^N with indices $k \geq k_1(\varepsilon)$, $N \geq N_\varepsilon$ we can follow through (2.9)–(2.12) to obtain (3.3) with $k_\varepsilon = k_1(\varepsilon) + l(\varepsilon) + 1$. \square

A traditional notion of mesh independence, described in [1] and [12], is that the iterate k after which the size of the projected gradient is less than ε is independent of N . To state a result of this type, we consider the difference between consecutive steps. This is feasible because the stepsizes α_k are assumed (see (2.3)) to be bounded from above and also uniformly from below away from zero. It holds that

$$k_N(\varepsilon) = \min\{k : \|u_k^N - u_{k-1}^N\|_H < \varepsilon\}, \quad k(\varepsilon) = \min\{k : \|u_k - u_{k-1}\|_H < \varepsilon\}.$$

The following mesh-independence result of the second type explains the results reported in Table 4.1.

THEOREM 3.2. *Assume that $u_k^N \rightarrow u_k$ in H for all $k \in \mathbb{N}$. Then for all $\varepsilon, \rho > 0$ there is $N_{\varepsilon, \rho}$ such that if $N \geq N_{\varepsilon, \rho}$, then*

$$k(\varepsilon + \rho) \leq k_N(\varepsilon) \leq k(\varepsilon).$$

Proof. By definition of $k(\varepsilon)$,

$$\nu(\varepsilon) := \varepsilon - \|u_{k(\varepsilon)} - u_{k(\varepsilon)-1}\|_H > 0.$$

The assumption of the theorem yields that for each $\mu > 0$ there are $N(\mu, k) \in \mathbb{N}$ such that

$$\|u_k^N - u_k\| < \mu \quad \text{for all } N \geq N(\mu, k).$$

In particular, we have that for

$$N > N_\varepsilon^1 := \max \{N(\nu(\varepsilon)/2, k(\varepsilon)), N(\nu(\varepsilon)/2, k(\varepsilon) - 1)\},$$

$$\begin{aligned} \|u_{k(\varepsilon)}^N - u_{k(\varepsilon)-1}^N\|_H &\leq \|u_{k(\varepsilon)} - u_{k(\varepsilon)-1}\|_H \\ &\quad + \|u_{k(\varepsilon)}^N - u_{k(\varepsilon)}\|_H + \|u_{k(\varepsilon)-1}^N - u_{k(\varepsilon)-1}\|_H \\ &< \varepsilon - \nu(\varepsilon) + \frac{1}{2}(\nu(\varepsilon) + \nu(\varepsilon)) \\ &= \varepsilon. \end{aligned}$$

Hence $k_N(\varepsilon) \leq k(\varepsilon)$.

On the other hand, for given $\varepsilon, \rho > 0$, we have

$$\|u_k - u_{k-1}\| \geq \varepsilon + \rho \quad \text{for all } k < k(\varepsilon + \rho).$$

Choose

$$N_{\varepsilon, \rho}^2 = \max \{N(\rho/2, k) : k = 1, \dots, k(\varepsilon + \rho)\}.$$

Then for all $N \geq N_{\varepsilon, \rho}^2$ and $k < k(\varepsilon + \rho)$

$$\begin{aligned} \|u_{k(\varepsilon)}^N - u_{k(\varepsilon)-1}^N\|_H &\geq \|u_{k(\varepsilon)} - u_{k(\varepsilon)-1}\|_H \\ &\quad - \|u_{k(\varepsilon)}^N - u_{k(\varepsilon)}\|_H - \|u_{k(\varepsilon)-1}^N - u_{k(\varepsilon)-1}\|_H \\ &\geq \varepsilon + \rho - \frac{1}{2}(\rho + \rho) \\ &= \varepsilon. \end{aligned}$$

Hence $k_N(\varepsilon) \geq k(\varepsilon + \rho)$ for all $N \geq N_{\varepsilon, \rho}^2$, and the statement follows with $N_{\varepsilon, \rho} = \max \{N_{\varepsilon}^1, N_{\varepsilon, \rho}^2\}$. \square

The third type of mesh-independence result considers the number of constraints that are changed in the transition from u_*^N to u_*^{N+1} . We will discuss this in the specific context of first-order differencing for the state and adjoint equations as described in §1. In this case,

$$(3.4) \quad \|s_*^N - s_*\|_\infty \leq \varepsilon_N.$$

In the case of an implicit Euler solution, $\varepsilon_N = ch_N$, where h_N is the stepsize at level N and the number of grid points is $1 + h_N^{-1}$. We accept $J^N = \{t \mid |s_*^N(t)| \geq \varepsilon_N\}$ as the active set at level n . Define $\Delta^N = J^N \Delta_{\mathcal{A}} J^{N+1}$. Here the A-symmetric difference of sets is

$$A \Delta_{\mathcal{A}} B = (A^c \cap B) \cup (B^c \cap A),$$

with $A^c = \mathcal{A} - (A \cap \mathcal{A})$ denoting the complement of A in \mathcal{A} . We state a set theoretic mesh-independence result.

THEOREM 3.3. *Let the assumptions of Theorem 3.1 hold. Assume that (2.14) and (3.4) hold and that $\varepsilon_N = \rho \varepsilon_{N-1}$ for some $\rho < 1$. Then there is \hat{c} independent of N , such that*

$$m(\Delta^N) \leq \hat{c} \varepsilon_N^{1/\mu}$$

for all N .

Proof. If $t \in (J^N)^c \cap J^{N+1}$, then

$$|s^*(t)| \leq |s_*^N(t)| + \varepsilon_N < 2\varepsilon_N$$

and, therefore, $t \in \mathcal{A} - J_{2\varepsilon_N}$. If $t \in (J^{N+1})^c \cap J^N$, then

$$|s^*(t)| \leq |s_*^{N+1}(t)| + \varepsilon_{N+1} \leq 2\varepsilon_{N+1} = 2\rho\varepsilon_N < 2\varepsilon_N.$$

Therefore $\Delta^N \subset \mathcal{A} - J_{2\varepsilon_N}$, which completes the proof by Lemma 2.2 with $\hat{c} = 2c$, where c is the constant in (2.15). \square

In the special case where \mathcal{Z} is a finite set, $\varepsilon_N = h_N = 2h_{N-1}$, and $\mu = 1$, the above result asserts that the number of grid points at level $N + 1$ for which

$$|s_*^N(t)| \geq \varepsilon_N, \quad (\text{approximately active at level } N)$$

but

$$|s_*^{N+1}(t)| < \varepsilon_{N+1}, \quad (\text{undetermined at level } N)$$

or vice versa, is bounded by a constant independent of N . To see this, note that the number of grid points in a set of size $\hat{c}h_N$ should be roughly \hat{c} . Hence only a finite number of constraint changes must be done if the initial iterate for level $N + 1$ is a converged result at level N .

If $\mu > 1$, however, the number of additional approximately active constraints that must be considered at each level will increase with N . The number of grid points in Δ_N will be approximately

$$m(\Delta_N)/\varepsilon_N = O(\varepsilon_N^{(1-\mu)/\mu}).$$

This estimate indicates that the number of changes for the number of approximately active constraints at each level is constant with $\mu = 1$ and increases for μ larger than 1.

To conclude this section, we show that the assumption on the uniform boundedness on α_k^N is satisfied for Armijo's rule under a uniform bound on the Lipschitz constant of F_N . The arguments are the same as for F and can be found in [2]. For completeness we include a proof.

LEMMA 3.4. *Let $u_k^N \in U_N$ be a sequence of points generated by the gradient projection algorithm with Armijo's stepsize rule (1.4). Assume that the constants $\tilde{\alpha}_N$ and δ_N specified for each N in (1.4) satisfy*

$$\tilde{\alpha}_N \geq \tilde{\alpha} > 0, \quad \delta_N \geq \delta > 0.$$

Suppose that there is $L > 0$ such that

$$\|\nabla F_N(u) - \nabla F_N(w)\| \leq L\|u - w\| \quad \text{for all } u, w \in U_N, \quad N \in \mathbb{N}.$$

Then there exists $\alpha \geq 0$ such that for all $k, N \in \mathbb{N}$,

$$\alpha_k^N \geq \alpha > 0$$

holds.

Proof. The definition of the projection implies that for all $u \in U_N$ and $\alpha > 0$

$$(3.5) \quad \langle \nabla F_N(u), u - P_N(u - \alpha \nabla F_N(u)) \rangle \geq \frac{1}{\alpha} \|u - P_N(u - \alpha \nabla F_N(u))\|^2$$

is true; see, e.g., [7, (2.6)]. The Lipschitz continuity yields for all $u, v \in D_N$ and all $N \in \mathbb{N}$

$$(3.6) \quad |F_N(u) - F_N(v) - \langle \nabla F_N(u), u - v \rangle| \leq \frac{1}{2}L\|u - v\|^2.$$

The Armijo stepsize rule implies that either $\alpha_k^N = \tilde{\alpha}_N$ or that with (3.5) and (3.6) the following inequalities hold:

$$\begin{aligned} & -\frac{L}{2} \|u_k^N - P_N(u_k^N - \frac{\alpha_N}{\beta_N} \nabla F_N(u_k^N))\|^2 \\ & \leq F_N(u_k^N) - F(P_N(u_k^N - \frac{\alpha_N}{\beta_N} \nabla F_N(u_k^N))) \\ & \quad - \langle \nabla F_N(u_k^N), u_k^N - P_N(u_k^N - \frac{\alpha_N}{\beta_N} \nabla F_N(u_k^N)) \rangle > \\ & \leq (\delta_N - 1) \langle \nabla F_N(u_k^N), u_k^N - P_N(u_k^N - \frac{\alpha_N}{\beta_N} \nabla F_N(u_k^N)) \rangle > \\ & \leq \frac{\delta_N - 1}{\alpha_N} \|u_k^N - P_N(u_k^N - \frac{\alpha_N}{\beta_N} \nabla F_N(u_k^N))\|^2, \end{aligned}$$

and therefore

$$2 \frac{1 - \delta}{L} \leq 2 \frac{1 - \delta_N}{L} \leq \alpha_k^N.$$

Hence we obtain

$$\alpha_k^N \geq \min\{\tilde{\alpha}, 2 \frac{1 - \delta}{L}\} > 0.$$

□

4. Numerical results. In this section we illustrate some of the theoretical results of the previous sections. Let us consider first a rather simple, but nevertheless illuminating, example from optimal control. The problem consists of minimizing

$$\frac{1}{2} \int_0^1 x(t)^2 dt,$$

subject to

$$\dot{x}(t) = b(t)u(t), \quad x(0) = 1, \quad t \in [0, 1]$$

and

$$u \in U = \{u \in L_2[0, 1] : |u(t)| \leq 0.5 \text{ a.e. in } [0, 1]\}.$$

Let us consider the following examples

Example 4.1. $b(t) = (t - 0.5)^3$;

Example 4.2. $b(t) = (t - 0.5)^3 + 0.03$;

Example 4.3. $b(t) = \cos(3t)$;

Example 4.4. $b(t) = (t - 0.5)|t - 0.5|$.

The first example does not satisfy the growth condition (1.11) for $\nu = 2$, but the second, which is a slightly perturbed example, does. As a more well-posed example, we use Example 4.3 where (1.11) is also satisfied.

The optimal control satisfies the following bang-bang-principle:

$$u_*(t) = -0.5 \operatorname{sgn} p_*(t)b(t) \quad \text{a.e. in } [0, 1],$$

where p_* solves the adjoint equation

$$-\dot{p}(t) = x_*(t), \quad p(1) = 0, \quad t \in [0, 1].$$

Since p is positive on $[0, 1)$ for all our choices of b , the function b determines the type of growth for the switching function

$$(4.1) \quad s_* = \frac{bp_*}{2}.$$

For the discrete case, we replace the feasible set U by

$$U_N = \{u \in L_2[0, 1] : u = \sum_{i=0}^{D_N-1} u_i \chi_i, |u_i| \leq 0.5 \ i = 0, \dots, D_N - 1\},$$

χ_i denoting the characteristic function on the interval $[i/(D_N), (i+1)/(D_N)]$, and the objective function F denoted by

$$(4.2) \quad F_N(u_N) = \frac{1}{2} \int_0^1 \left(\sum_{i=0}^{D_N-1} x_i \chi_i \right)^2 dt,$$

where x_i is the Euler approximation to the solution of the differential equation

$$x_{i+1} = x_i + \frac{1}{D_N} b_i u_i, \quad i = 0, \dots, D_N - 1, \quad x_0 = 1.$$

Here we set

$$b_i = b\left(\frac{i}{D_N}\right), \quad i = 0, \dots, D_N - 1.$$

The gradient of F_N can be computed by $s_N = \sum_{i=0}^{D_N-1} s_i \chi_i$, where

$$(4.3) \quad s_i = 0.5b_i p_i, \quad i = 0, \dots, D_N - 1.$$

Here p_i is the Lagrange multiplier that satisfies the discretized version of the adjoint differential equation by Euler's scheme.

To check the assumptions for Theorem 3.1, we note that all F_N and F are convex and continuously differentiable. Furthermore, the quadratic form of (4.2) implies under the use of the necessary optimality condition that

$$F_N(u^N) - F_N(u_*^N) \geq F_N(u^N - u_*^N) > 0$$

for all $u^N \in U_N$ with $u^N \neq u_*^N$. By invoking Theorem 1.2, we obtain uniform convergence of $F_N(u_k^N)$ to $F_N(u_*^N)$. More specifically, there is a constant $q \in \mathbb{R}$ with

$$(4.4) \quad \|x_k^N - x_*^N\|_{L_2} = F_N(u_k^N - u_*^N) \leq F_N(u_k^N) - F_N(u_*^N) \leq \frac{q}{k}$$

for all $k, N \in \mathbb{N}$. Since we know by inspection of the discretized problem that for sufficiently large N the optimal control u_*^N is bang-bang except for two subintervals, we have

$$\|u_* - u_*^N\|_{L_2} \rightarrow 0 \quad \text{and} \quad \|x_* - x_*^N\|_{L_\infty} \rightarrow 0.$$

From (4.1) and (4.3) we see that the convergence of the switching functions is determined by the corresponding adjoint states, which takes place in the L_∞ -norm because of the smoothing property of the adjoint differential and difference equation. Hence

$$\lim_{N, k \rightarrow \infty} \|s_k^N - s_*\|_{L_\infty} \leq \lim_{N, k \rightarrow \infty} \|s_k^N - s_*^N\|_{L_\infty} + \lim_{N \rightarrow \infty} \|s_*^N - s_*\|_{L_\infty} = 0.$$

Therefore, Theorem 3.1 is applicable to our examples, and we can expect mesh independence in the sense of the results in §3.

In Table 4.1 we do not terminate the algorithm on the identification of the active indices but rather when we reach a certain tolerance for the discrete L_2 -norm of the steps, i.e., the projected gradient:

$$k_N(10^{-4}) = \min\{k \in \mathbb{N} : \|u_k^N - u_{k-1}^N\| \leq 10^{-4}\}.$$

The table shows clearly that in all cases the number of iterations to reach this goal is independent on the dimension, which is in line with the results proved in Theorem 3.2.

We now discuss the condition $u_k^N \rightarrow u_k$ in H for all $k \in \mathbb{N}$, which occurs as an assumption in statement of Theorem 3.2. The iteration scheme (1.3) is given by

$$u_{k+1}^N = P(u_k^N - \alpha_k^N \nabla F_N(u_k^N)).$$

Our verification will be by induction and will be valid for almost every choice of the parameter ρ in the Armijo rule. This technical restriction on ρ is a consequence of the

TABLE 4.1
Termination based on small steps

D_N	k_N for Ex. 1	k_N for Ex. 2	k_N for Ex. 3
4	6	8	24
8	13	30	8
16	12	31	14
32	12	31	12
64	12	31	11
128	12	31	11
256	12	31	11

TABLE 4.2
Termination based on relative length of interval of active indices.

D_N	k^{90} for Ex. 1	k^{90} for Ex. 2	k^{90} for Ex. 3	k^{80} for Ex. 1
4	164	59	24	164
8	780	56	21	779
16	5032	76	14	734
32	4879	74	13	1065
64	9104	85	16	1048
128	9032	85	16	1040
256	8617	84	16	1069

discrete nature of the decision to accept or reject a step and was explained in detail in [11]. We refer the reader to that paper for a discussion of this point and will only note the issue below. Clearly, $u_0^N \rightarrow u_0$ uniformly. To complete the verification, we assume that $u_k^N \rightarrow u_k$ in L_2 then show that $u_{k+1}^N \rightarrow u_{k+1}$ in L_2 . Since $\nabla F_N(u_k^N) = s_k^N$, we can use similar arguments as above to check $s_k^N \rightarrow s_k$. The projection is continuous, and therefore $\alpha_k^N \rightarrow \alpha_k$ remains to be shown. This is true for almost every choice of ρ , as was the case in [11].

As a termination criterion in Table 4.2, we use the size of the interval where the indices are active. Termination occurs when the length of the interval corresponding to all active indices is a certain fixed percentage of the length of the total interval. Suppose we want to stop if p percent of the interval length is a bang-bang control; i.e., the corresponding indices are active. Then we define

$$k^p = \min\{k \in \mathbb{N} : |I(u_k)|/D_N \geq \frac{p}{100}\}$$

This termination criterion is related to Theorem 3.1 for an appropriate choice of ε . The results for this termination criterion are listed in Table 4.2 and show also the independence of the dimension of the problem. Example 4.1 is notably ill posed, so we also list the results for a less stringent termination criterion.

In Table 4.3 we illustrate the estimate in Theorem 3.3 on how the measure of set of almost active indices changes as the discretization is refined. Here the influence of the growth condition (2.14) can be documented for our examples. The values for ε_N in this table are $\varepsilon_N = .02h_N$ for Example 4.1, $\varepsilon_N = 10h_N$ for Example 4.3, and $\varepsilon = .1h_N$ for Example 4.4. According to the statement of Theorem 3.3 and the growth

TABLE 4.3
Change of active set through refinement.

Example 1			Example 3		Example 4	
D_N	$m(\Delta^N)$	$\frac{m(\Delta^N)}{m(\Delta^{N-1})}$	$m(\Delta^N)$	$\frac{m(\Delta^N)}{m(\Delta^{N-1})}$	$m(\Delta^N)$	$\frac{m(\Delta^N)}{m(\Delta^{N-1})}$
128	3.13'-2		3.05'-1		3.13'-2	
256	3.13'-2	1.00	1.45'-1	0.47	1.95'-2	0.63
512	2.34'-2	0.75	6.64'-2	0.46	1.76'-2	0.90
1024	1.76'-2	0.75	3.13'-2	0.47	1.27'-2	0.72
2048	1.42'-2	0.81	1.56'-2	0.50	8.30'-3	0.66
4096	1.12'-2	0.79	7.81'-3	0.50	5.86'-3	0.71
8192	8.91'-3	0.79	3.91'-3	0.50	4.15'-3	0.71

condition for our examples, we should expect an estimate for the ratios

$$m(\Delta^N)/m(\Delta^{N-1})$$

of the following type: It should be 0.79 for Example 4.1, 0.50 for Example 4.3, and 0.71 for Example 4.4. For high discretizations, these estimates can be verified in the given examples.

As a second example, we look at an optimal control problem with partial differential equations. For further references see, e.g., [13]. Let $y(t, x)$ denote the temperature at time $t \in [0, T]$ and at location $x \in [0, 1]$. Heat conduction problems with memory can be modeled with pseudoparabolic differential equations. The boundary problem is given by

$$\begin{aligned} y_t(t, x) &= y_{xx}(t, x) + \varepsilon y_{xtx}(t, x), & x \in (0, 1), \quad t \in (0, T), \\ y(0, x) &= 0, & x \in (0, 1), \\ y_x(t, 1) &= 0, & t \in (0, T), \\ -y_x(t, 0) - \varepsilon y_{xt}(t, 0) &= u(t), & t \in (0, T]. \end{aligned}$$

The function u denotes the input or control function. The variable ε represents a material constant. The objective function is given by

$$F(u) = \int_0^1 \phi(y(T, x) - z(x)) dx + \frac{\alpha}{2} \int_0^T u(t)^2 dt,$$

where $\phi: \mathbb{R} \rightarrow \mathbb{R}$ is a given continuously differentiable function and $\alpha \geq 0$ a given constant. The set of feasible controls is given by

$$U = \{u \in L_2[0, T] : |u(t)| \leq 1 \text{ a.e. in } [0, T]\}.$$

We define for $u \in L_2[0, T]$ the solution $y(T, \cdot)$, which enters the objective function as follows:

$$Su(x) := y(T, x) = \sum_{j=0}^{\infty} a_j^2 b_j c_j(x) \int_0^T e^{-\lambda_j(T-s)} u(s) ds$$

with the following abbreviations for $j = 0, 1, 2, \dots$:

$$\begin{aligned} c_j(x) &= \cos j\pi x, & x \in [0, 1] \\ \mu_k &= j^2 \pi^2 \\ b_j &= \frac{1}{(1+\varepsilon \mu_j)}, \\ \lambda_j &= \frac{\mu_j}{(1+\varepsilon \mu_j)}, \\ a_j &= \sqrt{2}, \quad k > 0 \quad a_0 = 1. \end{aligned}$$

TABLE 4.4
Pseudoparabolic Control Problem

D_N	$ I(u_{\text{iter}}) $	l_{act}	r_{act}	iter
10	9 (90%)	0.2000	0.2500	32
25	21 (84%)	0.2000	0.2800	31
50	44 (88%)	0.2100	0.2700	28
75	65 (87%)	0.2067	0.2733	29
100	86 (86%)	0.2050	0.2750	30
150	130 (87%)	0.2067	0.2733	29
200	173 (87%)	0.2050	0.2725	29

If we use the adjoint operator S^* of $S : L_2[0, T] \rightarrow C[0, 1]$, we can show that the optimal control u_* satisfies the following necessary optimality condition:

$$u_*(t) = -\text{sgn}(S^*(\phi'(Su(\cdot) - z(\cdot)))(t) + \alpha u_*(t)), t \in [0, T].$$

or, equivalently,

$$u_*(t) = -\text{sat}\left(\frac{1}{\alpha} S^*(\phi'(Su(\cdot) - z(\cdot)))(t)\right),$$

where the function sat is defined by (2.5).

We approximate the problem by truncating the infinite series representation of S by

$$S_N u(x) = \sum_{j=0}^{j_N} a_j^2 b_j c_j(x) \int_0^T e^{-\lambda_j(T-s)} u(s) ds;$$

we replace the set U by piecewise constant functions U_N on a fixed grid with values in $[-1, 1]$ as in control problem above. The integration, along the space variable x from 0 to 1, is replaced by a quadrature rule. Hence the function F_N is determined by

$$F_N(v) = \frac{1}{l_N + 1} \sum_{l=0}^{l_N} \phi\left(S_N\left(\sum_{i=0}^{D_N-1} v_i \chi_i\right)(x_l) - z(x_l)\right) + \frac{1}{N} \sum_{i=0}^{D_N-1} v_i^2.$$

We set

$$T = 0.5, \quad \varepsilon = 10^{-4}, \quad \alpha = 0.1, \quad j_N = 50, \quad l_N = 100, \quad \phi(y) = y^4, \quad z(x) = 3(x-0.5).$$

The termination criterion is set to be

$$\|u_{k+1}^N - u_k^N\| \leq 10^{-6}.$$

This is a reasonable termination criterion because the description of the optimal control u^* tells us that it is not a bang-bang control and has a singular arc. We tabulate in Table 4.4 the number of iterations iter needed to reach the termination criterion, the number of active indices $|I(u_{\text{iter}})|$, and the interval $[l_{\text{act}}, r_{\text{act}}]$, which was identified as singular; i.e., the control has values inside the interval $(-1, 1)$.

Acknowledgments. The authors thank the referees for helpful comments and remarks on the paper.

REFERENCES

- [1] E. L. ALLGOWER, K. BÖHMER, F. A. POTRA, AND W. C. RHEINOLDT, *A mesh-independence principle for operator equations and their discretizations*, SIAM J. Numer. Anal., 23 (1986), pp. 160–169.
- [2] D. B. BERTSEKAS, *On the Goldstein–Levitin–Polyak gradient projection method*, IEEE Trans. Automat. Control, (1976), pp. 174–184.
- [3] ———, *Projected Newton methods for optimization problems with simple constraints*, SIAM J. Control Optim., 20 (1982), pp. 221–246.
- [4] P. H. CALAMAI AND J. MORÉ, *Projected gradient methods for linearly constrained problems*, Math. Programming, 39 (1987), pp. 93–116.
- [5] A. R. CONN, N. I. M. GOULD, AND P. L. TOINT, *Global convergence of a class of trust region algorithms for optimization problems with simple bounds*, SIAM J. Numer. Anal., 25 (1988), pp. 433–460.
- [6] A. L. DONTCHEV, *Perturbations, Approximations and Sensitivity Analysis of Optimal Control Systems*, Vol. 52, Lecture Notes in Control and Inform. Sci., Springer–Verlag.
- [7] J. C. DUNN, *Global and asymptotic convergence rate estimates for a class of projected gradient processes*, SIAM J. Control Optim., 19 (1981), pp. 368–400.
- [8] E. M. GAFNI AND D. P. BERTSEKAS, *Two-metric projection methods for constrained optimization*, SIAM J. Control Optim., 22 (1984), pp. 936–964.
- [9] M. GAWANDE AND J. C. DUNN, *Variable metric gradient projection processes in a convex feasible set defined by nonlinear inequalities*, J. Appl. Math. Optim., 17 (1988), pp. 103–119.
- [10] A. A. GOLDSTEIN, *On Newton's method*, Numer. Math., 7 (1965), pp. 391–393.
- [11] C. T. KELLEY AND E. W. SACHS, *Mesh independence of Newton-like methods for infinite dimensional problems*, J. Integral Equations. Appl. to appear.
- [12] ———, *Broyden's method for approximate solution of nonlinear integral equations*, J. Integral Equations Appl., 9 (1985), pp. 25–44.
- [13] ———, *Applications of Quasi-Newton Methods to Pseudoparabolic Control Problems*, in Optimal Control of Partial Differential Equations II—Theory and Applications, May 1986, Birkhäuser, Basel 1987.
- [14] E. S. LEVITIN AND B. T. POLYAK, *Constrained optimization methods*, USSR Comput. Math. Phys., 6 (1966), pp. 1–50.
- [15] J. MORÉ, *Trust regions and projected gradients*, Tech. Report ANL/MCS-TM-107, Argonne National Laboratory, Math. and Comp. Science Div. Report, 1988.
- [16] P. L. TOINT, *Global convergence of a class of trust-region methods for nonconvex minimization in Hilbert space*, IMA J. Numer. Anal., 8 (1988), pp. 231–252.

CONTRÔLE OPTIMAL DANS LES ÉQUATIONS ELLIPTIQUES*

R. TAHRAOUI†

Abstract. The goal of this work is to study problems of control when the state is governed by an elliptic partial differential equation. The control dependency is nonconvex. These kinds of problems arise in such applications as domain control or hydrodynamics. Examples are provided.

Key words. contrôle optimal, équations elliptiques

1. Introduction et notation. On se propose, dans ce travail, de traiter quelques questions d'existence de contrôle optimal dans un cadre non convexe, et où l'état du système est défini par des équations uniformément elliptiques.

Il existe de nombreux travaux traitant de ce sujet. Les premiers résultats semblent remonter à 1955 avec le travail de Krein [17] qui, motivé par l'étude des zones de stabilité de l'équation

$$y'' + \lambda p(x)y = 0$$

où $x \in]-\infty, +\infty[$ et p est une fonction période, s'était intéressé aux deux problèmes suivants:

$$\inf \{ \lambda_n(v) / v \in \mathcal{U} \},$$

$$\max \{ \lambda_n(v) / v \in \mathcal{U} \},$$

où $\lambda_n = \lambda_n(v)$ est la n ième valeur propre de

$$y'' + \lambda_n \cdot v \cdot y = 0,$$

$$y(0) = y(1) = 0;$$

le contrôle v appartient à l'ensemble \mathcal{U} donné par

$$\mathcal{U} = \left\{ v \in L^\infty(0, 1) / h \leq v \leq H, \int_0^1 v \, dx = M \right\},$$

h, H, M étant des constantes positives données. Les résultats obtenus sont très précis (cf. [17]: la dimension d'espace $n = 1$ en est une des raisons). Motivé par l'étude des vibrations d'une barre Banks [18] aborda les deux questions précédentes pour l'opérateur du quatrième ordre suivant:

$$y^{(4)} - \lambda \cdot v y = 0, \quad y(0) = y(1) = y''(0) = y''(1) = 0,$$

le contrôle v appartenant à \mathcal{U} . Nous pouvons remarquer que, aussi bien dans [17] que dans [18], le problème de l'existence ne se pose pas: la dépendance du problème en fonction du contrôle est convexe.

* Received by the editors September 9, 1988; accepted for publication (in revised form) November 30, 1990.

† Centre de Recherches de Mathématiques de la Décision, Université de Paris-Dauphine, Place de Maréchal de Lattre Tassigny, 75775 Paris Cedex 16, France.

Dans un cadre plus général que dans [17] et [18], Zérah [19] donne des résultats d'existence, pour des situations non convexes, en utilisant diverses techniques, notamment l'homogénéisation et les réarrangements, pour certains problèmes de la forme suivante:

$$\begin{aligned}
 & -(a(v)y')' = \lambda_1 \cdot b(v) \cdot y \\
 (\mathcal{P}_1) \quad & J(v) = \lambda_1(v), \text{ qui est non convexe} \\
 & \sup \{J(v)\},
 \end{aligned}$$

sous différentes conditions aux limites, le contrôle v appartenant à un ensemble de type \mathcal{U} . De plus, Zérah donne plusieurs motivations mécaniques du problème (\mathcal{P}_1) dans la recherche des formes optimales:

- (i) recherche de la colonne la plus résistante,
- (ii) recherche de la colonne la plus haute, à volume donné (Tour Eiffel),
- (iii) vibrations transversales d'une poutre,
- (iv) minimisation de la déviation d d'une poutre dont le poids est donné i.e. minimiser:

$$J(a) = - \int_0^1 \sin \theta(s) ds$$

sous les contraintes:

$$\begin{aligned}
 & \frac{d}{ds} \left(a^2 \frac{d\theta}{ds} \right) = \cos \theta \\
 & \theta(0) = \theta'(1) = 0, \quad \int_0^1 a(s) ds = 1.
 \end{aligned}$$

On remarquera que J n'est pas convexe.

Signalons également que dans [26] on trouvera d'autres problèmes de contrôle non convexe comme, par exemple, le problème suivant: (anisotropie optimale d'une barre en torsion) il s'agit de maximiser la fonctionnelle

$$K(\alpha) = 2 \int_{\Omega} \varphi(x, y) dx dy,$$

où φ désigne l'état solution de l'équation suivante:

$$\begin{aligned}
 & -\operatorname{div} (A(\alpha) \cdot \nabla \varphi) = 2 \quad \text{dans } \Omega; \\
 & \varphi/\Gamma = 0, \quad \text{où } \Gamma = \partial\Omega
 \end{aligned}$$

dans laquelle α est une fonction contrôle décrivant l'orientation; la dépendance en α de $A(\cdot)$ est non convexe. Les motivations précédentes constituent une première justification de ce travail. Enfin signalons un résultat très significatif de [19]: (\mathcal{P}_1) possède au moins une solution si la fonction $v \rightarrow a(v) \cdot b(v)$ est strictement monotone. Ce résultat intéressant semble lié à la dimension d'espace $n = 1$.

En effet en dimension d'espace $n \geq 2$, d'une manière générale, si on se place dans une situation où la dépendance du contrôle est non convexe, l'existence d'une solution pour des problèmes du type (\mathcal{P}_1) n'est pas toujours assurée. Il est connu (cf. [28]) que le problème de contrôle de domaine suivant:

$$\begin{aligned}
 & -\operatorname{div} (v \cdot \nabla \omega) = 1 \\
 (\mathcal{P}_2) \quad & \omega/\Gamma = 0 \\
 & \inf (J(v), / v \in \mathcal{U}_1\}, \quad J(v) = \int_{\Omega} \omega dx = \int_{\Omega} v \cdot |\nabla \omega|^2 dx
 \end{aligned}$$

n'a pas de solutions si l'on prend

$$\mathcal{U}_1 = \{v \in L^\infty(\Omega) / v(x) \in \{\alpha, \beta\}, \text{mes}(v^{-1}(\alpha)) = \theta_1, \text{mes}(v^{-1}(\beta)) = \theta_2, \theta_1 + \theta_2 = \text{mes}(\Omega)\}.$$

Il se produit un phénomène d'homogénéisation: la solution optimale $\bar{\omega}$ exige un mélange de matériaux, i.e., le contrôle optimal \bar{u} vérifie

$$\alpha \leq \bar{u}(x) \leq \beta \quad \text{p.p. } x \in \Omega$$

avec

$$\int_{\Omega} \bar{u} \, dx = \text{mes}(\Omega), \quad \text{mes}(\{x \in \Omega / \alpha < \bar{u}(x) < \beta\}) > 0.$$

Dans notre travail le contrôle intervient dans les termes d'ordre zéro de l'opérateur. Il s'agit, par exemple, de maximiser l'énergie

$$J(v) = \int_{\Omega} |\nabla \omega|^2 \, dx,$$

où ω vérifie l'équation d'état:

$$(\mathcal{P}_3) \quad -\Delta \omega = v, \quad \omega / \Gamma = 0,$$

v étant le contrôle admissible qui appartient à l'ensemble

$$\mathcal{U}_2 = \{v \in L^\infty(\Omega) / \alpha \leq v \leq \beta, v_* = u_0\}^1,$$

où v_* est le réarrangement croissant unidimensionnel de v ; il est clair que \mathcal{U}_2 est non convexe et est une généralisation de \mathcal{U}_1 : il s'agit de mélange à un nombre, éventuellement, infini de matériaux.

Deux types d'applications importantes sont modélisés par (\mathcal{P}_3) .

(1) *Contrôle de domaine.* Dans ce cas $u_0(x) \in \{\alpha, \beta\}$ p.p. $x \in \Omega$; l'idée est de relaxer le problème au sens suivant, i.e., résoudre

$$(\mathcal{PP}_3) \quad \inf \{J(v) / v \in \mathcal{U}_3\},$$

quand l'ensemble des contrôles admissibles est

$$\mathcal{U}_3 = \left\{ v \in L^\infty(\Omega) / \alpha \leq v \leq \beta, \int_{\Omega} v \, dx = \gamma \right\}.$$

Ce problème relaxé admet une solution optimale $(\bar{u}, \bar{\omega})$. On conclut que le problème initial admet une solution en montrant que le contrôle \bar{u} est bang-bang. Cette dernière propriété importante sera, pour cette raison, étudiée dans un cadre assez général. Cette formulation englobe une classe assez large de problèmes d'optimisation de forme, i.e., de contrôle par le domaine tels que le problème de la rigidité à la torsion d'une barre de section droite de forme quelconque ou le problème de la capacité: [30].

(2) *Problème d'hydrodynamique.* Dans ce cas u_0 est une fonction croissante à valeurs dans \mathbb{R}^+ (signalons au passage le cas intéressant où cardinal de $(\text{Im } u_0)$ est fini); le problème (\mathcal{P}_3) modélise, entre autres, un problème d'hydrodynamique où ω représente le potentiel des vitesses $(\partial \omega / \partial x_2, -\partial \omega / \partial x_1)$, $J(v) = \int_{\Omega} |\nabla \omega|^2 \, dx$ représente l'énergie cinétique du fluide et $-\Delta \omega$ la vorticit  dont la fonction de distribution est fix e. Pour la formulation de ce probl me nous renvoy r   [36], [29], et [31].

¹ Ce type de contrainte se rencontre dans [12] dont nous avons re u le preprint en d cembre 1985 au moment o  la r daction de ce travail  tait achev e.

Tous ces problèmes entrent dans une formulation générale pour laquelle nous donnerons, dans un premier temps, des résultats d'existence sous des hypothèses assez simples; et en deuxième lieu nous fournirons, autant que possible, des informations qualitatives sur les solutions optimales. L'unicité sera également abordée. Enfin signalons que ce travail a été annoncé dans [15] et une première version en a été donnée dans [16].

Etant donné un ouvert borné régulier de R^n , on considère l'ensemble des contrôles admissibles

$$(1.1) \quad \mathcal{U}(\alpha, \beta, \gamma) = \mathcal{U} = \left\{ v \in L^\infty(\Omega), 0 < \alpha(x) \leq v(x) \leq \beta(x) \text{ p.p. } x, \int_{\Omega} v \, dx = \gamma \right\},$$

où α et β sont des fonctions de $L^\infty(\Omega)$ satisfaisant

$$0 < \alpha(x) < \beta(x) \quad \text{p.p. } x \in \Omega.$$

On suppose \mathcal{U} non vide, i.e., que γ vérifie

$$\int_{\Omega} \alpha \, dx \leq \gamma \leq \int_{\Omega} \beta \, dx.$$

Pour tout v dans \mathcal{U} , l'équation d'état du système est

$$(1.2) \quad \begin{aligned} -\Delta \omega &= a(x, v, \omega) \quad \text{dans } \Omega, \\ \omega/\Gamma &= 0, \end{aligned}$$

dans laquelle la fonction a de $\Omega \times \mathbb{R}^+ \times \mathbb{R}$ dans \mathbb{R} , supposée de Carathéodory satisfait la condition

$$(1.3) \quad a(x, s, t) \geq a_0 \quad \text{p.p. } x \in \Omega, \quad \forall (s, t) \in \mathbb{R}^+ \times \mathbb{R}.$$

Suivant les besoins du contexte, nous noterons la fonction état du système par ω , $\omega(v)$, $\omega(a)$, ou $\omega(a, v)$ si aucune ambiguïté n'est à craindre; et à (1.2) on associe la fonction coût

$$(1.4) \quad J(v) = \int_{\Omega} l(v) \cdot b(x, \omega) \, dx + \int_{\Omega} c(x, \omega) \, dx,$$

pour laquelle les fonctions b, c sont de caratheodory, l une fonction continue et ω une solution de (1.2), associée au contrôle v .

DÉFINITION 1.1. Le problème $P = P(a, b, c, l)$ consiste à trouver u dans \mathcal{U} tel que

$$J(u) = \inf \{J(v), v \in \mathcal{U}\}.$$

2. Resultats preliminaires. Nous avons besoin des hypothèses suivantes:

$$(2.1) \quad v \rightarrow a(x, v, t) \quad \text{est convexe, p.p. } x \in \Omega, \quad \forall t \in \mathbb{R}$$

et satisfait

$$(2.2) \quad \begin{aligned} |a(x, s, t)| &\leq r_1 |t| + r_2(x) \\ \text{p.p. } x \in \Omega, \quad \forall t \in \mathbb{R}, \quad \forall s \in [\inf \alpha, \sup \beta], \end{aligned}$$

où r_1 est une constante positive satisfaisant $r_1 < c(\Omega)$ constante de Poincaré; r_2 est une

fonction de $L^2(\Omega)$.

$$(2.3) \quad b(x, t) \geq 0 \quad \text{p.p. } x \in \Omega, \quad \forall t \in \mathbb{R}$$

$$(2.4) \quad b(x, t) \text{ croissante en } t, \text{ p.p. } x \in \Omega$$

$$(2.5) \quad c(x, t) \text{ croissante en } t, \quad \text{p.p. } x \in \Omega$$

$$(2.6) \quad l(t) \text{ est convexe, positive.}$$

$$(2.7) \quad a(x, s, t) \text{ est croissante en } t, \quad \text{p.p. } x, \quad \forall s.$$

Remarque 2.1. L'hypothèse (2.2) est faite pour simplifier la présentation et éviter de poser le problème de l'existence d'une solution de (1.2) qui n'est pas dans le but que nous nous sommes fixé ici.

Nous avons le résultat suivant.

THÉORÈME 2.1. *Sous les hypothèses (1.3) et (2.1) à (2.6) le problème $P(a, b, c, l)$ admet au moins une solution optimale $(\bar{u}, \bar{\omega} = \omega(\bar{u}))$.*

Démonstration. L'idée de la preuve, basée essentiellement sur le principe du maximum, a déjà servi pour un énoncé voisin mais pour une équation linéaire [8]. Sous (2.2), pour tout v dans \mathcal{U} , l'existence d'une solution de (1.2) est classique. Soit v_n une suite minimisante de $J(\cdot)$; on a

$$(2.8) \quad \begin{aligned} -\Delta \omega_n &= a(x, v_n, \omega_n) \quad \text{dans } \Omega, \\ \omega_n / \Gamma &= 0, \end{aligned}$$

et

$$J(v_n) \rightarrow \inf \{J(v), v \in \mathcal{U}\} \quad \text{quand } n \rightarrow +\infty;$$

la fonction ω_n étant borné dans $H_0^1(\Omega)$, il existe une sous-suite notée encore ω_n telle que

$$\begin{aligned} v_n &\rightarrow \bar{u} \quad \text{dans } L^\infty(\Omega) \text{ faible}^* \\ \omega_n &\rightarrow \tilde{\omega} \quad \text{dans } H_0^1(\Omega) \text{ faible} \\ \omega_n &\rightarrow \tilde{\omega} \quad \text{dans } L^2(\Omega) \text{ fort et p.p. } x \in \Omega. \end{aligned}$$

Et nous avons

$$(2.9) \quad \inf J(v) = \lim J(v_n) \geq \int_{\Omega} l(\bar{u}) b(x, \tilde{\omega}) dx + \int_{\Omega} c(x, \tilde{\omega}) dx.$$

Il s'agit de montrer que

$$J(\bar{u}) = \inf \{J(v), v \in \mathcal{U}\};$$

pour cela il suffit de montrer qu'il existe $z = \omega(\bar{u})$ solution de (1.2) tel que $\tilde{\omega} > z$. Un passage à la limite dans (2.8) donne, à une sous-suite près

$$\begin{aligned} -\Delta \tilde{\omega} &= \tilde{g}(x) \quad \text{dans } \Omega \\ \tilde{\omega} / \Gamma &= 0, \end{aligned}$$

où \tilde{g} est la limite faible dans $L^2(\Omega)$ de $\tilde{g}_n(x) = a(x, v_n(x), \omega_n(x))$, et qui vérifie:

$$\tilde{g}(x) \geq g(x) = a(x, \bar{u}(x), \tilde{\omega}(x)).$$

Nous allons définir une suite z_n de fonctions de $H_0^1(\Omega)$ dont la limite, à une sous-suite près, sera un état optimal z associé au contrôle \bar{u} . Pour cela considérons les équations récurrentes suivantes:

$$\begin{aligned} z_0 &= \tilde{\omega} \\ -\Delta z_n &= a(x, \bar{u}, z_{n-1}) \quad \text{dans } \Omega, \quad n \geq 1 \\ z_{n/\Gamma} &= 0. \end{aligned}$$

Le principe du maximum [13], la croissance de $a(x, s, \cdot)$ et (1.3) nous permettent d'affirmer que l'on a

$$(2.10) \quad z_1 \geq z_2 \geq \dots \geq z_n \geq y_0 \quad \text{dans } \Omega,$$

où y_0 est solution de: $-\Delta y_0 = a_0$, $y_{0/\partial\Omega} = 0$; ces inégalités (2.10) permettent d'établir que z_n est borné dans $H_0^1(\Omega)$; par conséquent, à une sous-suite près, z_n converge vers z dans $H_0^1(\Omega)$ faible et presque partout; z vérifie l'équation

$$(2.11) \quad -\Delta z = a(x, \bar{u}(x), z(x)), \quad z_{/\Gamma} = 0$$

et l'inégalité

$$(2.12) \quad \tilde{\omega}(x) \geq z(x) \quad \text{dans } \Omega.$$

Enfin la croissance de b et c donnent

$$\int_{\Omega} l(\bar{u})b(x, \tilde{\omega}) \, dx + \int_{\Omega} c(x, \tilde{\omega}) \, dx \geq \int_{\Omega} l(\bar{u})b(z, x) \, dx + \int_{\Omega} c(x, z) \, dx = J(\bar{u});$$

et cette inégalité jointe à (2.9) entraîne

$$\inf J(v) \geq J(\bar{u});$$

i.e.,

$$\inf J(v) = J(\bar{u});$$

par conséquent $(\bar{u}, z = \omega(\bar{u}))$ est solution optimale du problème. \square

Abordons maintenant l'unicité de la solution. Nous avons le théorème suivant.

THÉOREME 2.2. *On suppose que les fonctions $a(x, s, t)$ et $c(x, t)$ sont strictement croissantes par rapport à t et $a(x, s, t)$ strictement convexe en s . Alors sous les hypothèses du Théorème 2.1, nous avons les résultats suivants:*

(i) *si (u_1, ω_1) et (u_2, ω_2) désignent deux solutions optimales, alors on a $\omega_1 \leq \omega_2$ dans Ω ;*

(ii) *il y a unicité de la solution optimale si $b \equiv 0$.*

Démonstration.

Première étape. Soit deux solutions (u_1, ω_1) , (u_2, ω_2) de $P(a, b, c, l)$. Nous allons montrer que l'on ne peut avoir

$$|\{x \in \Omega / \omega_1(x) < \omega_2(x)\}| > 0 \quad \text{et} \quad |\{x \in \Omega / \omega_2(x) < \omega_1(x)\}| > 0.$$

En effet, dans le cas contraire, posons $z_0 = \inf(\omega_1, \omega_2)$ et considérons la fonction z_n solution de

$$-\Delta z_n = a(x, u_1, r_{n-1}), \quad z_{n/\partial\Omega} = 0,$$

où

$$r_0 = z_0, \quad r_n = \inf(z_n, z_{n-1}).$$

La croissance stricte de $a(x, s, \cdot)$ et le principe du maximum fort entraînent que

$$\omega_1 > z_1 \geq z_2 \geq \cdots \geq z_n \geq y_0 \quad \text{dans } \Omega$$

(la première inégalité s'entend pour au moins une composante connexe de Ω). A une sous-suite près, z_n et r_n convergent, dans $H_0^1(\Omega)$ fort, vers z solution de

$$-\Delta z = a(x, u_1, z)$$

$$z/\partial\Omega = 0,$$

et satisfaisant $z < \omega_1$. Ceci entraîne la contradiction suivante:

$$\int_{\Omega} l(u_1)b(x, z) \, dx + \int_{\Omega} c(x, z) \, dx < \inf \{J(v), v \in \mathcal{U}\};$$

ainsi $\omega_1 \leq \omega_2$ dans Ω .

Deuxième étape. Supposons $b(x, t) \equiv 0$; d'après la première étape nous avons $\omega_1(x) \leq \omega_2(x)$ pour tout x . De plus, si l'ensemble $\{x \in \Omega / \omega_1(x) < \omega_2(x)\}$ est de mesure non nulle, nous aurons

$$J(u_1) = \int_{\Omega} c(x, \omega_1) \, dx < \int_{\Omega} c(x, \omega_2) \, dx = J(u_2)$$

grâce à la stricte croissance de $c(x, \cdot)$. Ce qui contredit le fait que u_2 est optimal. Donc nous avons bien $\omega_1 = \omega_2$. Enfin, montrons que $u_1 = u_2$. Pour cela supposons $E = \{x / u_1(x) \neq u_2(x)\}$ de mesure > 0 . Par la stricte convexité de $a(x, \cdot, t)$ et l'égalité $\omega_1 = \omega_2$, nous avons

$$\begin{aligned} -\Delta \omega_1 &= \frac{1}{2} a(x, u_1, \omega_1) + \frac{1}{2} a(x, u_2, \omega_1) \\ &> a\left(x, \frac{u_1 + u_2}{2}, \omega_1\right) \quad \text{p.p. } x \in E. \end{aligned}$$

Considérons alors la suite \tilde{y}_n de fonctions, solution de l'équation

$$-\Delta \tilde{y}_n = a(x, \tilde{u}, \tilde{y}_{n-1}), \quad y_{n/\partial\Omega} = 0,$$

où

$$\tilde{u} = \frac{u_1 + u_2}{2}, \quad \tilde{y}_0 = \omega_1.$$

Par le principe du maximum fort cette suite vérifie

$$\omega_1 > \tilde{y}_1 \geq \tilde{y}_2 \geq \cdots \geq \tilde{y}_n \geq y_0 \quad \text{dans } \Omega.$$

A une sous-suite près, $\tilde{y}_n \rightarrow \tilde{y} = \tilde{y}(\tilde{u})$ solution de (1.2) avec $v = \tilde{u}$. On montre, grâce à la stricte croissance de $c(x, \cdot)$, que

$$J(\tilde{u}) = \int_{\Omega} c(x, \tilde{y}) \, dx < \inf \{J(v), v \in \mathcal{U}\} = J(u_1);$$

ce qui constitue une contradiction. Donc $u_1 = u_2$.

Remarque 2.2. (1) La stricte croissance de $c(x, \cdot)$ au voisinage de $t = 0$ suffit.

(2) Pour simplifier, nous supposons dans toute la suite α et β constants.

3. Dependance non convexe du contrôle. La situation la plus générale consiste à envisager le cas où ni l ni h ne sont convexes. Cependant pour simplifier la présentation de la démonstration, il nous a paru utile d'envisager séparément les trois cas suivants:

- (i) l est affine;
- (ii) h est affine;
- (iii) la situation générale.

3.1. On suppose que l est affine positive croissante sur $[\alpha, \beta]$, i.e.,

$$(3.1) \quad l(t) = l_0 \cdot t + s_0 \geq 0 \quad \forall t \in [\alpha, \beta], \quad l_0 > 0;$$

on désigne par $x = (x_1, x_2, \dots, x_n)$ un élément générique de Ω ; et pour tout i dans $\{1, 2, \dots, n\}$ on note $x = (x_i^v, x_i)$ où $x_i^v = (x_i, \dots, x_{i-1}, x_{i+1}, x_n)$. Nous supposons $c(x, t)$ régulière, $b(x, t)$ régulière telle que

(3.2) pour tout i dans $\{1, 2, \dots, n\}$ l'application $(x_i, t) \rightarrow b(x_i^v, x_i, t)$ est concave, et la fonction a de la forme

$$(3.3) \quad a(x, s, t) = f(x) + h(s)$$

avec h une fonction définie sur $[\alpha, \beta]$ régulière et f une fonction de $L^\infty(\Omega)$. Posons

$$K(h) = K = \{s \in \mathbb{R}^+ / h^{**}(s) < h(s)\};$$

on a

$$K = \bigcup_{i \in I} K_i, \quad K_i = K_i(h),$$

où I est un ensemble au plus dénombrable et K_i est un intervalle de \mathbb{R}^+ pour tout i . Il existe deux constantes μ_i et ν_i telles que

$$h^{**}(s) = \mu_i \cdot s + \nu_i \quad \forall s \in K_i.$$

Enfin on fait l'hypothèse suivante

$$(3.4) \quad \frac{\partial b}{\partial t}(x, t) \cdot \frac{\partial}{\partial s} [l(s) \cdot (f(x) + h^{**}(s))] + \frac{\partial c}{\partial t}(x, t) \cdot \frac{\partial}{\partial s} (f(x) + h^{**}(s)) > 0$$

p.p. $x \in \Omega$, $\forall s \in K(h)$, $\forall t \in \mathbb{R}$.

Remarque 3.1. Si $a(x, s, t) \geq 0$, on peut supposer, grâce au principe du maximum, (3.4) vrai seulement pour $t \in \mathbb{R}^+$. Ceci est également valable pour toute la suite.

Nous avons le résultat suivant.

THÉORÈME 3.1. *Sous les hypothèses (2.3) à (2.5), (3.1) à (3.4), le problème $P(a, b, c, l)$ admet au moins une solution optimale $(\bar{u}, \bar{\omega} = \omega(\bar{u}))$.*

Démonstration. On relaxe le problème (P) , i.e., on remplace, dans l'équation d'état, $a(x, s, t)$ par $f(x) + h^{**}(s)$ sa convexifiée en s ; on note le problème ainsi obtenu par (P^{**}) : $P^{**} = P(a^{**}, b, c, l)$. D'après le Théorème 2.1., (P^{**}) admet au moins une solution optimale $(\bar{u}, \bar{\omega} = \omega(\bar{u}))$. On se propose alors de démontrer que

$$h^{**}(\bar{u}(x)) = h(\bar{u}(x)) \quad \text{p.p. } x \in \Omega \text{ (étape 1),}$$

ensuite que $(\bar{u}, \bar{\omega})$ est aussi solution de (P) (étape 2).

Etape 1. Elle repose sur la relation d'extrémalité suivante dont la preuve est classique

$$(3.5) \quad \int_{\Omega} l'(\bar{u}) b(x, \bar{\omega}) \delta v \cdot dx + \int_{\Omega} \left[l(\bar{u}) \frac{\partial b}{\partial t}(x, \bar{\omega}) + \frac{\partial c}{\partial t}(x, \bar{\omega}) \right] \delta \omega \cdot dx \geq 0$$

pour tout δv accroissement admissible du contrôle, avec la fonction $\delta\omega$ solution de

$$(3.6) \quad -\Delta(\delta\omega) = \frac{dh^{**}}{ds}(\bar{u}) \cdot \delta v, \quad \delta\omega/\Gamma = 0.$$

Pour éliminer $\delta\omega$ dans (3.5) on introduit l'état adjoint

$$(3.7) \quad -\Delta\bar{p} = l(\bar{u}) \frac{\partial b}{\partial t}(x, \bar{\omega}) + \frac{\partial c}{\partial t}(x, \bar{\omega}), \quad \bar{p}/\Gamma = 0.$$

On multiplie (3.6) par \bar{p} et (3.7) par $\delta\omega$; on intègre par parties, ensuite on retranche membre à membre les deux équations obtenues; le résultat obtenu injecté dans (3.5) donne

$$(3.8) \quad \int_{\Omega} [l'(\bar{u})b(x, \bar{\omega}) + h^{**'}(\bar{u}) \cdot \bar{p}] \delta v \cdot dx \geq 0$$

pour tout δv accroissement admissible du contrôle.

PROPOSITION 3.1. *Sous l'hypothèse (3.2) vérifiée par $b(x, t)$, la fonction $b(x, \bar{\omega})$ appartient à $H_{\text{loc}}^2(\Omega)$ et satisfait*

$$(3.9) \quad -\Delta(b(x, \bar{\omega})) \geq \frac{\partial b}{\partial t}(x, \bar{\omega}) \cdot (f(x) + h^{**}(\bar{u})) \text{ dans } \Omega.$$

Démonstration. La fonction $b(x, \bar{\omega})$ appartient à $H^1(\Omega) \cap H_{\text{loc}}^2(\Omega)$: soit une suite $\omega_k \in H_0^1$, régulière convergeant vers $\bar{\omega}$ dans $H_0^1(\Omega) \cap W^{2,4}(\Omega) \cap C(\bar{\Omega})$; nous avons

$$(3.10) \quad -\Delta(b(x, \omega_k)) = - \sum_{i=1}^n E_i(x, \omega_k) - \frac{\partial b}{\partial t}(x, \omega_k) \cdot \Delta\omega_k,$$

où la fonction E_i a pour expression

$$E_i(x, \omega_k) = \frac{\partial^2 b}{\partial t^2}(x, \omega_k) \cdot \left(\frac{\partial \omega_k}{\partial x_i} \right)^2 + 2 \frac{\partial^2 b}{\partial x_i \partial t}(x, \omega_k) \cdot \frac{\partial \omega_k}{\partial x_i} + \frac{\partial^2 b}{\partial x_i^2}(x, \omega_k);$$

par l'hypothèse (3.2) $E_i(x, \omega_k)$ est négative; de plus elle est bornée dans $L^2(\Omega)$. A une sous-suite près, on passe à la limite dans (3.10)

$$-\Delta b(x, \bar{\omega}) = - \frac{\partial b}{\partial t}(x, \bar{\omega}) \cdot \Delta\bar{\omega} + \xi \quad \text{p.p. } x \in \Omega$$

où la fonction ξ est la limite dans $L^2(\Omega)$ faible de $-\sum_{i=1}^n E_i(x, \bar{\omega}_k)$; i.e., en utilisant l'équation d'état de (P^{**})

$$(3.11) \quad -\Delta b(x, \bar{\omega}) = \frac{\partial b}{\partial t}(x, \bar{\omega}) \cdot (f(x) + h^{**}(\bar{u})) + \xi;$$

ce qui entraîne que $b(x, \bar{\omega})$ appartient à $H_{\text{loc}}^2(\Omega)$ et

$$-\Delta b(x, \bar{\omega}) \geq \frac{\partial b}{\partial t}(x, \bar{\omega}) \cdot (f(x) + h^{**}(\bar{u})) \quad \text{dans } \Omega,$$

puisque ξ est positive par l'hypothèse (3.2). \square

Nous sommes maintenant en mesure de démontrer le résultat annoncé pour cette étape.

PROPOSITION 3.2. *On a*

$$h^{**}(\bar{u}(x)) = h(\bar{u}(x)) \quad p.p. \ x \in \Omega.$$

Démonstration. Il suffit de montrer que l'ensemble

$$F = \{x \in \Omega / h^{**}(\bar{u}(x)) < h(\bar{u}(x))\}$$

est de mesure nulle. Si l'on pose

$$F_i = \{x \in F / u(x) \in K_i\}$$

on a

$$F = \bigcup_{i \in I} F_i.$$

Supposons alors qu'il existe $i_0 \in I$ tel que $\text{mes}(F_{i_0}) > 0$; et considérons la fonction

$$S(x) = l_0 \cdot b(x, \bar{\omega}) + \mu_{i_0} \cdot \bar{p}(x) \quad \text{dans } \Omega;$$

elle appartient à $H_{\text{loc}}^2(\Omega)$ d'après (3.7) et la Proposition 3.1.; de plus elle satisfait $-\Delta S > 0$ dans F_{i_0} par (3.4). Ainsi S n'a pas de palier sur F_{i_0} ; ceci entraîne que

$$(3.12a) \quad \text{mes}(\{x \in F_{i_0} / S(x) = t\}) = 0 \quad \forall t \in \mathbb{R}.$$

A l'aide de cette propriété de S , nous allons montrer que $\text{mes}(F_{i_0}) > 0$ est en contradiction avec la condition nécessaire d'optimalité (3.8); pour cela on se donne v dans

$$\mathcal{U}_{i_0} = \left\{ v \in L^\infty(F_{i_0}) / \alpha \leq v \leq \beta, \int_{F_{i_0}} v \, dx = \int_{F_{i_0}} \bar{u} \, dx \right\};$$

le contrôle admissible

$$\tilde{v} = \begin{cases} \bar{u} & \text{sur } \Omega \setminus F_{i_0} \\ v & \text{sur } F_{i_0}, \end{cases}$$

porté dans (3.8) permet d'avoir:

$$(3.12b) \quad \int_{F_{i_0}} S(x)(v - \bar{u}) \, dx \geq 0 \quad \forall v \in \mathcal{U}_{i_0}.$$

On peut supposer par exemple $F_{i_0}^+ = \{x \in F_{i_0} / S(x) \geq 0\}$ de mesure non nulle et connexe, quitte à travailler sur une composante connexe de mesure non nulle; posons

$$G(t) = \{x \in F_{i_0}^+ / S(x) \geq t\} \quad \text{pour } t \geq 0$$

et définissons

$$R(t) = \text{mes}(G(t)) \quad \forall t \geq 0;$$

c'est une fonction continue par (3.12a). Il existe donc $t_0 > 0$ tel que

$$R(t_0) = \text{mes}(G(t_0)) = \frac{k_1 \text{mes}(F_{i_0}^+) - \int_{F_{i_0}^+} u \, dx}{k_1 - k_0},$$

où k_1 et k_0 , ($k_1 > k_0$), représentent les extrémités de K_{i_0} .

Dans le premier membre de (3.12b) prenons le "contrôle admissible" suivant

$$v_0 = \begin{cases} \bar{u} & \text{sur } F_{i_0}^- = F_{i_0} \setminus F_{i_0}^+ \\ k_0 & \text{sur } G(t_0) \\ k_1 & \text{sur } F_{i_0}^+ \setminus G(t_0); \end{cases}$$

et majorons cette intégrale; nous obtenons

$$\int_{F_{i_0}} S(x)(v_0 - \bar{u}) \, dx < t_0 \int_{G(t_0)} (k_0 - \bar{u}) \, dx + t_0 \int_{F_{i_0 \setminus G(t_0)}^+} (k_1 - \bar{u}) \, dx = 0;$$

ceci contredit (3.12b). Donc mes $(F_{i_0}^+) = 0$; de même $F_{i_0}^- = \{x \in F_{i_0} / S(x) < 0\}$ est de mesure nulle; et ainsi s'achève la preuve de la Proposition 3.2 et donc fin de l'étape 1. \square

Etape 2. $(\bar{u}, \bar{\omega})$ est aussi solution de $P(a, b, c, l)$. En effet, la solution $(\bar{u}, \bar{\omega})$ de (P^{**}) satisfait l'équation d'état de (P) :

$$-\Delta \bar{\omega} = f(x) + h(\bar{u}), \quad \bar{\omega}|_{\Gamma} = 0$$

puisque $h(\bar{u}) = h^{**}(\bar{u})$; si l'on note par $\omega_{**} = \omega_{**}(v)$ la solution de l'équation d'état de (P^{**}) correspondant à v , on a

$$J(\bar{u}) \leq \int_{\Omega} l(v)b(x, \omega_{**}) \, dx + \int_{\Omega} c(x, \omega_{**}) \, dx;$$

comme par le principe du maximum $\omega(v) \geq \omega_{**}$, il s'en suit grâce à la croissance de b et c

$$J(\bar{u}) \leq J(v) \quad \forall v \in \mathcal{U}. \quad \square$$

3.2. Enfin pour les deux cas qui restent nous supposons, à la place de (3.1) l'hypothèse

(3.1a) l^{**} est positive et croissante sur $K(l)$.

De plus, nous supposons dans le cas (ii)

(3.13) h est affine.

$$(3.14) \quad \frac{\partial b}{\partial t}(x, t) \cdot \frac{\partial}{\partial s} [l^{**}(s) \cdot (f(x) + h(s))] + \frac{\partial c}{\partial t}(x, t) \cdot \frac{\partial}{\partial s} (f(x) + h(s)) > 0$$

p.p. $x \in \Omega$, $\forall t \in \mathbb{R}$, $\forall s \in K(l)$;

dans le cas (iii)

$$(3.15) \quad K(h) \equiv K(l)$$

$$(3.16) \quad \frac{\partial b}{\partial t}(x, t) \cdot \frac{\partial}{\partial s} [l^{**}(s) \cdot (f(x) + h^{**}(s))] + \frac{\partial c}{\partial t}(x, t) \cdot \frac{\partial}{\partial s} (f(x) + h^{**}(s)) > 0$$

p.p. $x \in \Omega$, $\forall t \in \mathbb{R}$, $\forall s \in K(h) = K(l)$.

Nous avons alors le corollaire suivant.

COROLLAIRE 3.1. *On fait les hypothèses (2.3) à (2.5), (3.2), (3.3) et (3.1a). Alors le problème $P(a, b, c, l)$ admet au moins une solution dans les deux situations suivantes:*

(1) (3.13) et (3.14) ont lieu;

(2) (3.15) et (3.16) ont lieu.

La preuve de ce résultat est identique, dans ses grandes lignes, à celle du Théorème 3.1: si $(\bar{u}, \bar{\omega})$ désigne une solution du problème relâché $P(a^{**}, b, c, l^{**})$ alors la fonction $S = l^{**}(\bar{u}) \cdot b(x, \bar{\omega}) + h^{**}(\bar{u}) \cdot \bar{p}$ est ici également sans palier sur l'ensemble $F = \{x \in \Omega / \bar{u}(x) \in K(l)\}$. La suite de la démonstration demeure inchangée par rapport à celle du Théorème 3.1.

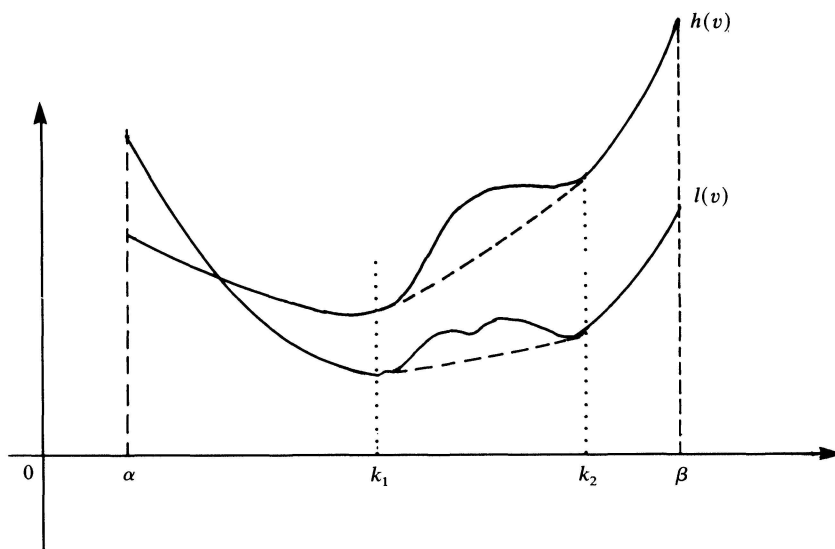


FIG. 1

Exemple 3.1.

$$\begin{aligned}
 b(x, t) &= t, \quad c(x, t) \equiv 0, \quad f(x) \equiv 0, \\
 K(h) &= K(l) =]k_1, k_2[\\
 -\Delta \omega &= h(v) \quad \text{dans } \Omega, \quad h(v) \geq 0 \quad \text{sur } [\alpha, \beta], \\
 \omega/\Gamma &= 0, \\
 J(v) &= \int_{\Omega} l(v) \cdot \omega \, dx;
 \end{aligned}$$

(cf. Fig. 1). La condition de positivité (3.16) devient:

$$\frac{d}{ds} (l^{**}(s) \cdot h^{**}(s)) > 0 \quad \text{sur }]k_1, k_2[.$$

Et si par exemple $h = l > 0$, la condition ci-dessus se réduit à $dl^{**}/ds(s) > 0$ sur $]k_1, k_2[$.

Remarque 3.2. Le Théorème 3.1 demeure valide si l'on suppose (3.4) < 0 a condition de modifier convenablement les autres hypothèses pour avoir $-\Delta S < 0$, i.e., S sans palier.

4. Contrôle bang-bang. Dans cette partie on suppose toujours α et β constants; et on fait les hypothèses suivantes:

$$\begin{aligned}
 (4.1) \quad h(v) &\geq H(v) = \theta_1 \cdot v + \delta_1, \quad \forall v \in [\alpha, \beta] \\
 h(\alpha) &= H(\alpha), \quad h(\beta) = H(\beta),
 \end{aligned}$$

$$\begin{aligned}
 (4.2) \quad l(v) &\geq L(v) = \theta_2 \cdot v + \delta_2, \quad \forall v \in [\alpha, \beta] \\
 l(\alpha) &= L(\alpha), \quad l(\beta) = L(\beta), \quad \theta_2 > 0
 \end{aligned}$$

$$\begin{aligned}
 (4.3) \quad \frac{\partial b}{\partial t}(x, t) \cdot \frac{\partial}{\partial s} [L(s) \cdot (f(x) + H(s))] &+ \frac{\partial c}{\partial t}(x, t) \cdot \frac{\partial}{\partial s} (f(x) + H(s)) > 0 \\
 \text{p.p. } x \in \Omega, \quad \forall t \in \mathbb{R}, \quad \forall s \in [\alpha, \beta].
 \end{aligned}$$

Nous avons le résultat suivant.

THÉORÈME 4.1. *On suppose les hypothèses (4.1)–(4.3), (2.3)–(2.5), (3.2), et (3.3). Alors le problème $P(a, b, c, l)$ possède au moins une solution optimale $(\bar{u}, \bar{\omega} = \omega(\bar{u}))$. De plus cette solution possède la propriété bang-bang; de façon plus précise si l'on pose*

$$S(x) = \theta_2 \cdot b(x, \bar{\omega}) + \theta_1 \cdot \bar{p}(x) \quad \forall x \in \Omega,$$

\bar{p} étant l'état adjoint associé à $\bar{\omega}$, il existe $t > 0$ tel que le contrôle optimal \bar{u} s'écrit

$$(4.3a) \quad \bar{u} = \begin{cases} \alpha & \text{sur } \Omega_\alpha \\ \beta & \text{sur } \Omega_\beta = \Omega \setminus \Omega_\alpha \end{cases}$$

où

$$\Omega_\alpha = \{x \in \Omega / S(x) \geq t\}.$$

Démonstration. Le problème relaxé $P(a^{**}, b, c, l^{**})$ possède au moins une solution optimale $(\bar{u}, \bar{\omega} = \omega(\bar{u}))$. Comme dans la preuve du Théorème 3.1 on montre que la fonction $S = \theta_2 \cdot b(x, \bar{\omega}) + \theta_1 \cdot \bar{p}$ est sans palier sur tout Ω ; ce qui entraîne que \bar{u} est de la forme (4.3a) en vertu de la condition d'optimalité

$$\int_{\Omega} S(x)(v - \bar{u}) \, dx \geq 0 \quad \forall v \in \mathcal{U}.$$

La propriété bang-bang de \bar{u} et les hypothèses (4.1) et (4.2) entraînent que $(\bar{u}, \bar{\omega})$ est solution optimale de $P(a, b, c, l)$ en vertu du principe du maximum et de la croissance de b et c .

Remarque 4.1. Le résultat précédent reste valable si l'on remplace \mathcal{U} donné en (1.1) par l'ensemble non convexe

$$\mathcal{U}' = \left\{ v \in L^\infty(\Omega) / \alpha \leq v \leq \beta, \int_{\Omega} g(v) \, dx = \gamma \right\}$$

où g est, par exemple, une fonction convexe, pour simplifier, strictement monotone, après quelques modifications naturelles en (4.3).

Pour illustrer cette remarque donnons un exemple.

Exemple 4.1.

$$-\Delta \omega = v$$

$$\omega|_{\partial\Omega} = 0 \quad \text{dans } \Omega, \quad J(v) = \int_{\Omega} v \cdot \omega \, dx$$

où v appartient

$$\mathcal{U} = \left\{ v \in L^\infty(\Omega) / v(x) \in \{\alpha, \beta\} \text{ p.p. } x \in \Omega \text{ et } \int_{\Omega} k(v) \, dx = \gamma \right\}$$

qu'on suppose non vide; k est une fonction strictement croissante de $[\alpha, \beta]$ dans \mathbb{R}^+ telle que son graphe soit en-dessous de la corde joignant les deux points $(\alpha, k(\alpha))$ et $(\beta, k(\beta))$. On fait un changement de fonction contrôle $V = k(v)$. Le problème devient:

$$-\Delta \omega = k^{-1}(V) \quad \text{dans } \Omega,$$

$$\omega|_{\partial\Omega} = 0, \quad J(V) = \int_{\Omega} k^{-1}(V) \cdot \omega \, dx,$$

$$\mathcal{U}' = \left\{ V \in L^\infty(\Omega) / V(x) \in \{k(\alpha), k(\beta)\} \text{ p.p. } x \in \Omega, \text{ et } \int_{\Omega} V \, dx = \gamma \right\}.$$

Il est alors aisé de vérifier que les conditions du Théorème 4.1 sont réunies.

Le résultat du Théorème 4.1 nous amène à la question naturelle: les hypothèses (4.1) et (4.2) sont-elles nécessaires pour que $P(a, b, c, l)$ admette un contrôle optimal bang-bang?

La réponse sera positive si l'on modifie quelque peu la condition de positivité (3.16): on la suppose vraie pour tout $s \in [\alpha, \beta]$, i.e.,

$$(4.4a) \quad \frac{\partial b}{\partial t}(x, t) \cdot \frac{\partial}{\partial s} [l^{**}(s) \cdot (f(x) + h^{**}(s))] + \frac{\partial c}{\partial t}(x, t) \cdot \frac{\partial}{\partial s} (f(x) + h^{**}(s)) > 0 \\ \text{p.p. } x \in \Omega, \quad \forall t \in \mathbb{R}, \quad \forall s \in [\alpha, \beta];$$

si aucune des deux fonctions l et h n'est affine, on suppose que l'on a de plus

$$(4.4b) \quad K(h) \equiv K(l).$$

THÉORÈME 4.2. *On suppose que la fonction b est strictement positive et on fait les hypothèses (2.4), (2.5), (3.2), (3.3), et (4.4). Alors, (4.1) et (4.2) forment une condition nécessaire et suffisante d'existence d'un contrôle optimal bang-bang.*

Démonstration. (i) la condition suffisante résulte du Théorème 4.1.

(ii) Pour montrer que (4.1) et (4.2) sont nécessaires, nous raisonnerons par l'absurde. D'après le Corollaire (3.1) le problème $P(a, b, c, l)$ admet au moins une solution $(\bar{u}, \bar{\omega})$; supposons, par exemple, que (4.2) ne soit pas vrai, i.e.,

$$(4.5) \quad l^{**'}(\alpha) < l^{**'}(\beta)$$

et que \bar{u} soit bang-bang; $(\bar{u}, \bar{\omega})$ satisfait la condition d'optimalité

$$\int_{\Omega} S(x)(v - \bar{u}) dx \geq 0 \quad \forall v \in \mathcal{U},$$

où

$$S = l^{**'}(\bar{u}) \cdot b(x, \bar{\omega}) + h^{**'}(\bar{u}) \cdot \bar{p}.$$

Cette condition peut aussi s'écrire

$$(4.6) \quad S(x) \cdot v + \mu v \geq S(x)\bar{u}(x) + \mu \cdot \bar{u}(x) \\ \text{p.p. } x \in \Omega, \quad \forall v \in [\alpha, \beta]$$

où μ est un multiplicateur de Lagrange. Mais puisque $S(x)$ est sans palier par (4.4), on peut montrer qu'il existe t_0 tel que

$$\bar{u}(x) = \begin{cases} \alpha & \text{sur } \Omega_{\alpha} \\ \beta & \text{sur } \Omega_{\beta} \end{cases}$$

où

$$\Omega_{\alpha} = \{x \in \Omega / S(x) > t_0\}$$

$$\Omega_{\beta} = \{x \in \Omega / S(x) \leq t_0\}$$

et par conséquent (4.6) entraîne les deux inégalités

$$(4.7) \quad l^{**'}(\alpha)b(x, \bar{\omega}(x)) + \theta \bar{p}(x) + \mu \geq 0 \quad \text{p.p. } x \in \Omega_{\alpha},$$

$$(4.8) \quad l^{**'}(\beta)b(x, \bar{\omega}(x)) + \theta \bar{p}(x) + \mu \leq 0 \quad \text{p.p. } x \in \Omega_{\beta} = \Omega \setminus \Omega_{\alpha}$$

où $\theta = h^{**'}(\alpha) = h^{**'}(\beta)$. A l'aide de (4.5) majorons l'inégalité (4.7) et minorons (4.8), en utilisant la stricte positivité de b :

$$\forall \delta \in [l^{**'}(\alpha), l^{**'}(\beta)], \quad \delta \cdot b(x, \bar{\omega}(x)) + \theta \cdot \bar{p}(x) + \mu \geq 0 \quad \text{dans } \Omega_{\alpha}$$

$$\forall \delta \in [l^{**'}(\alpha), l^{**'}(\beta)], \quad \delta \cdot b(x, \bar{\omega}(x)) + \theta \cdot \bar{p}(x) + \mu \leq 0 \quad \text{dans } \Omega_{\beta}.$$

Ceci entraîne que l'une des deux fonctions $\bar{\omega}$ et \bar{p} n'est pas continue; d'où la contradiction. Si (4.1) n'a pas lieu ou si (4.1) et (4.2) n'ont pas lieu le raisonnement demeure inchangé. \square

Exemple 4.2.

$$-\Delta\omega = v \quad \text{dans } \Omega,$$

$$\omega|_{\partial\Omega} = 0, \quad J(v) = \int_{\Omega} l(v) \cdot \omega \, dx + \int_{\Omega} \omega^2 \, dx,$$

où $l(v)$ est une fonction convexe positive strictement croissante sur $[\alpha, \beta]$. Il est clair que dans ce cas la condition de positivité (4.4.a) est satisfaite. L'ensemble admissible \mathcal{U} étant toujours donné par (1.1), on suppose γ tel que

$$\int_{\Omega} \alpha \, dx < \gamma < \int_{\Omega} \beta \, dx.$$

D'après le théorème (3.1) il existe \bar{u} appartenant à \mathcal{U} tel que

$$J(\bar{u}) \leq J(v) \quad \forall v \in \mathcal{U}.$$

Le Théorème 4.2 entraîne que

$$|\{x \in \Omega / \alpha < \bar{u}(x) < \beta\}| > 0;$$

i.e., que la solution \bar{u} n'est pas bang-bang.

Par contre si l'on supprime, dans \mathcal{U} , la condition sur la masse $\int_{\Omega} v \, dx = \gamma$, le problème

$$\inf [J(v) / v \in L^{\infty}(\Omega), v(x) \in [\alpha, \beta] \text{ p.p.}]$$

admet l'unique solution $u_0 = \alpha$; et ce problème est identique au suivant

$$\inf [J(v) / v(x) \in \{\alpha, \beta\} \text{ p.p.}].$$

Cet exemple montre que la condition de masse fixée $\int_{\Omega} v \, dx = \gamma$, absente dans [3], joue ici un rôle important dans l'existence (cf. Théorème 3.1) ou la non existence d'un contrôle optimal bang-bang. C'est sa présence qui nécessite les différentes conditions de positivité de type (3.4), (4.4), etc. Dans [3] si l'on introduit la condition de masse, les résultats obtenus ne semblent pas s'adapter (tout au moins de façon immédiate).

5. Un ensemble de contrôle plus général. Dans ce paragraphe on choisit α et β constants, $a(x, s, t) = f(x) + h(s)$ qu'on suppose positive pour simplifier, et h et l affines. Pour la commodité du lecteur rappelons, pour toute fonction k positive mesurable de Ω dans \mathbb{R} , les définitions classiques suivantes [9], [10].

DÉFINITIONS 5.1. (1) Pour tout mesurable E de Ω , $|E|$ désigne la mesure de Lebesgue de E .

(2) La fonction de distribution de k est la fonction δ_k de \mathbb{R} dans $[0, |\Omega|]$, définie par

$$\delta_k(t) = |\{x \in \Omega / k(x) < t\}|.$$

(3) Le réarrangement croissant (unidimensionnel) de k est la fonction notée k_* de $[0, |\Omega|]$ dans \mathbb{R} telle que

$$(5.1) \quad \begin{aligned} k_*(s) &= \inf \{t \in \mathbb{R} / \delta_k(t) > s\} \text{ si } s \in [0, |\Omega|[, \\ k_*(|\Omega|) &= \sup_{\Omega} \text{ess } k. \end{aligned}$$

(4) Le réarrangement décroissant (unidimensionnel) de k est défini par

$$(5.2) \quad k^*(s) = k_*(|\Omega| - s) \text{ p.p. } s.$$

(5) On appelle ensemble de niveau ou ensemble équipotentiel de k tout sous-ensemble de Ω du type

$$E_1(t) = \{x \in \Omega / k(x) > t\}, \quad t \in \mathbb{R}$$

ou

$$E_2(t) = \{x \in \Omega / k(x) < t\}, \quad t \in \mathbb{R}$$

avec éventuellement des inégalités au sens large. Et nous désignerons par “courbes ou lignes de niveau” de k les ensembles tels que

$$E_3(t) = \{x \in \Omega / k(x) = t\}, \quad t \in \mathbb{R}.$$

(6) k est dit sans palier si $|\{x \in \Omega / k(x) = t\}| = 0 \quad \forall t \in \mathbb{R}$. Etant donnée une fonction croissante u_0 définie sur $[0, |\Omega|]$, telle que

$$0 < \alpha \leq u_0(s) \leq \beta \quad \text{p.p. } s \in [0, |\Omega|],$$

on introduit, à la place de (1.1), l'ensemble des contrôles admissibles suivant:

$$(5.3) \quad \mathcal{U} = \{v \in L^\infty(\Omega) / \alpha \leq v \leq \beta, v_* = u_0\}.$$

Et on se propose de résoudre le problème $P(a, b, c, l)$ qui, rappelons-le, consiste à trouver \bar{u} dans \mathcal{U} satisfaisant

$$J(\bar{u}) = \inf \{J(v), v \in \mathcal{U}\},$$

où

$$J(v) = \int_{\Omega} l(v)b(x, \omega) dx + \int_{\Omega} c(x, \omega) dx$$

et ω la fonction état solution de

$$-\Delta \omega = f(x) + h(v), \quad \omega / \partial \Omega = 0.$$

Nous avons le résultat suivant.

THÉORÈME 5.1. *On suppose réalisées les conditions suivantes: (2.3), (3.2), (4.4a) et (5.3), h affine croissante et l affine croissante positive. Alors le problème $P(a, b, c, l)$ possède une solution optimale $(\bar{u}, \bar{\omega} = \omega(\bar{u}))$. De plus cette solution possède les propriétés suivantes:*

(1) $\bar{\omega}$ et \bar{u} ont “les mêmes ensembles équipotentiels” au sens suivant: tout $t > 0$ il existe $s = s(t)$ tel que

$$\{x \in \Omega / \bar{u}(x) < s\} \subseteq \{x \in \Omega / l'(\bar{u})b(x, \bar{\omega}) + h'(\bar{u})\bar{p} \geq t\} \subseteq \{x \in \Omega / \bar{u}(x) \leq s\},$$

où \bar{p} désigne l'état adjoint associé à $\bar{\omega}$.

(2) \bar{u} est fonction décroissante de $l'(\bar{u})b(x, \bar{\omega}) + h'(\bar{u})\bar{p}$; de façon plus précise \bar{u} satisfait la relation

$$\bar{u} = \phi(l'(\bar{u})b(x, \bar{\omega}) + h'(\bar{u})\bar{p})$$

où

$$\phi(t) = u_0([l'(\bar{u})b(x, \bar{\omega}) + h'(\bar{u})\bar{p}]^*)^{-1}(t).$$

Démonstration. L'ensemble \mathcal{U} n'étant pas convexe, on se propose de relaxer le problème (P) en remplaçant \mathcal{U} par

$$\begin{aligned} \tilde{\mathcal{U}} = \left\{ v \in L^\infty(\Omega) / \alpha \leq v \leq \beta, \int_0^t v_*(s) ds \geq \int_0^t u_0(s) ds, \right. \\ \left. \forall t \in [0, |\Omega|], \text{ et } \int_0^{|\Omega|} v_*(s) ds = \int_0^{|\Omega|} u_0(s) ds \right\}; \end{aligned}$$

cet ensemble est convexe puisqu'il s'écrit:

$$\tilde{\mathcal{U}} = \left\{ v \in L^\infty(\Omega) / \alpha \leq v \leq \beta, \int_E v(x) dx \geq \int_0^{|E|} u_0(s) ds, \right. \\ \left. \forall \text{ mesurable } E \subseteq \Omega, \text{ et } \int_\Omega v(x) dx = \int_0^{|\Omega|} u_0(s) ds \right\}.$$

Le problème ainsi obtenu sera noté (\tilde{P}) . D'après le résultat de Migliaccio [11],² $\tilde{\mathcal{U}}$ est la fermeture de \mathcal{U} pour la topologie $\sigma(L^\infty, L^1)$. Donc (\tilde{P}) possède une solution $(\bar{u}, \bar{\omega})$; et on se propose de montrer que $(\bar{u}, \bar{\omega})$ est aussi solution de (P) , i.e., que \bar{u} appartient à \mathcal{U} .

Pour cela, nous partons de la relation d'extrémalité

$$(5.4) \quad \int_\Omega [l'(\bar{u})b(x, \bar{\omega}) + h'(\bar{u})\bar{p}](v - \bar{u}) dx \geq 0, \quad \forall v \in \tilde{\mathcal{U}},$$

où \bar{p} , défini par (3.7), représente l'état adjoint associé à $\bar{\omega}$; et on applique l'inégalité de Hardy-Littlewood:

$$(5.5) \quad \int_\Omega S(x)v(x) dx \geq \int_\Omega S(x)\bar{u}(x) dx \geq \int_0^{|\Omega|} S^*(s) \cdot \bar{u}_*(s) ds,$$

où, pour alléger les notations, nous avons posé

$$S = l'(\bar{u})b(x, \bar{\omega}) + h'(\bar{u})\bar{p}.$$

L'inégalité

$$A(t) = \int_0^t \bar{u}_*(s) ds \geq \int_0^t u_0(s) ds = A_0(t)$$

nous permet de minorer le dernier membre de (5.5) en effectuant deux intégrations par parties:

$$\int_0^{|\Omega|} S^*(s) \cdot \bar{u}_*(s) ds = S^*(|\Omega|) \cdot A(|\Omega|) - \int_0^{|\Omega|} \frac{dS^*(s)}{ds} A(s) ds \\ \geq S^*(|\Omega|) \cdot A_0(|\Omega|) - \int_0^{|\Omega|} \frac{dS^*(s)}{ds} A_0(s) ds$$

puisque $A(|\Omega|) = A_0(|\Omega|)$; d'où, après la deuxième intégration par parties:

$$(5.6) \quad \int_0^{|\Omega|} S^*(s) \cdot \bar{u}_*(s) ds \geq \int_0^{|\Omega|} S^*(s) \cdot u_0(s) ds.$$

D'après la proposition (3.1) et la condition de positivité (4.4a), la fonction $S(x)$ est continue sans palier; ce qui permet d'introduire

$$\phi(t) = u_0((S^*)^{-1})(t), \quad \psi(t) = t \cdot \phi(t);$$

cette définition de ϕ et l'équimesurabilité de $\psi(S(x))$ et $\psi(S^*(s))$ permettent de transformer (5.6) en

$$(5.7) \quad \int_0^{|\Omega|} S^*(s) \cdot \bar{u}_*(s) ds \geq \int_0^{|\Omega|} S^*(s) \cdot u_0(s) ds = \int_0^{|\Omega|} \psi(S^*(s)) ds \\ = \int_\Omega \psi(S(x)) dx.$$

² On peut également se reporter à [12].

Et on compare (5.7) avec les inégalités (5.5) dans lesquelles on remplace v par le contrôle admissible $\phi(S(x))$; on obtient la série d'égalités:

$$\begin{aligned} \int_{\Omega} S(x) \cdot \phi(S(x)) \, dx &= \int_{\Omega} S(x) \cdot \bar{u}(x) \, dx = \int_0^{|\Omega|} S^*(s) \cdot \bar{u}_*(s) \, ds \\ &= \int_0^{|\Omega|} S^*(s) \cdot u_0(s) \, ds. \end{aligned}$$

La dernière inégalité entraîne que $\bar{u}_* = u_0$ et la deuxième que S et \bar{u} ont les "mêmes ensembles équipotentiels," i.e., pour tout $t > 0$ il existe $s = s(t)$ tel que

$$\{x \in \Omega / \bar{u}(x) < s\} \subseteq \{x \in \Omega / S(x) \geq t\} \subseteq \{x \in \Omega / \bar{u}(x) \leq s\}.$$

Et enfin il est clair que $\bar{u}(x) = \phi(S(x))$ avec ϕ décroissante. \square

Abordons maintenant la question de l'unicité. Signalons tout d'abord qu'elle reste posée dans le cas général; et que nous ne disposons que d'une réponse très partielle donnée par le théorème suivant.

THÉORÈME 5.2. *On suppose les hypothèses du Théorème 5.1. Alors si l'ensemble des contrôles optimaux de (\tilde{P}) est convexe, (P) admet une solution unique.*

Démonstration. Une conséquence de la preuve du Théorème 5.1 est que toute solution de (P) est aussi solution de (\tilde{P}) , i.e., que les deux problèmes sont identiques; soit u_1 et u_2 deux contrôles optimaux de (\tilde{P}) ; par hypothèse $\theta u_1 + (1 - \theta)u_2$ est aussi contrôle optimal de (\tilde{P}) pour tout $\theta \in]0, 1[$; donc $\theta u_1 + (1 - \theta)u_2$ appartenant à \mathcal{U} pour tout θ , i.e.,

$$(\theta u_1 + (1 - \theta)u_2)_* = u_0.$$

Ceci entraîne $u_1 = u_2$. En effet, soit k une fonction régulière strictement convexe; on a par équimesurabilité

$$\int_{\Omega} k(\theta u_1 + (1 - \theta)u_2) \, dx = \int_0^{|\Omega|} k(u_0) \, ds \quad \forall \theta \in]0, 1[;$$

d'où

$$\frac{d}{d\theta} \int_{\Omega} k(\theta u_1 + (1 - \theta)u_2) \, dx = 0;$$

ce qui donne

$$\int_{\Omega} (k'(u_1) - k'(u_2))(u_1 - u_2) \, dx = 0,$$

i.e., $u_1 = u_2$. \square

Exemple 5.1 (Unicité). $f = 0$, $h(v) = v$, $b(x, t) = 0$, et $c(x, t)$ strictement croissante et convexe en t à x fixé:

$$-\Delta \omega = v, \quad \omega|_{\partial\Omega} = 0,$$

$$J(v) = \int_{\Omega} c(x, \omega) \, dx.$$

Exemple 5.2 (Contrôle de domaine). Etant données trois constantes $0 < \alpha = \alpha_1 < \alpha_2 < \alpha_3 = \beta$, on considère l'ensemble des contrôles admissibles suivant:

$$U = \left\{ D = (D_1, D_2, D_3), \text{ mesurable } \subset \Omega^3 / |D_i| = \gamma_i, D_i \cap D_j = \emptyset \text{ si } i \neq j, \bigcup_{i=1,3} D_i = \Omega \right\};$$

et on se pose le problème

$$(\pi) \quad \begin{aligned} -\Delta \omega_D &= v_D, & \omega_{D/\partial\Omega} &= 0, \\ \inf \{J(D), D \in U\}, \end{aligned}$$

où

$$v_D(x) = \alpha_i \quad \text{si } x \in D_i \quad i = 1, 2, 3,$$

et

$$J(D) = \int_{\Omega} v_D \cdot \omega_D dx.$$

Ce problème est équivalent à (P) puisque pour tout $D \in U$, $(v_D)_* = u_0$ qui est défini par

$$u_0(s) = \begin{cases} \alpha_1 & \text{si } 0 \leq s < \gamma_1, \\ \alpha_2 & \text{si } \gamma_1 \leq s < \gamma_1 + \gamma_2, \\ \alpha_3 & \text{si } \gamma_1 + \gamma_2 \leq s < \gamma_1 + \gamma_2 + \gamma_3 = |\Omega|. \end{cases}$$

Par conséquent, il existe $t_1 > 0$ et $t_2 > 0$ tels que (π) admet une solution $D^0 = (D_1^0, D_2^0, D_3^0)$ caractérisé par

$$D_1^0 = \{x \in \Omega / \omega_{D^0}(x) > t_1\},$$

$$D_2^0 = \{x \in \Omega / t_1 \leq \omega_{D^0}(x) > t_2\},$$

$$D_3^0 = \{x \in \Omega / t_2 \leq \omega_{D^0}(x)\}.$$

6. Applications.

6.1. Rigidité maximale à la torsion. On considère une poutre P creuse. (Cf. Fig. 2) Le creux est représenté par Ω_0 un ouvert fixé. L'aire de la section droite Σ de P étant fixée égale à β , il s'agit de trouver un contour déterminant Σ pour que la rigidité de P soit maximale si elle était soumise à des forces de torsion. Il s'agit là d'un problème typique d'optimisation de forme. Nous allons montrer que l'on peut appliquer

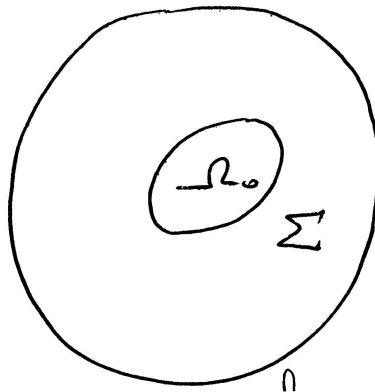


FIG. 2

les idées du § 4 que nous avons mises en oeuvre dans un cadre général. Ce problème a été considéré par divers auteurs; citons par exemple [26], [21], [20]. Et très récemment une étude en est faite avec une approche différente de la nôtre dans [20].

Nous supposons que la section droite du matériau occupe l'ouvert Σ et que le matériau est homogène, isotrope. Nous renvoyons à ([23], [26]) pour la présentation du modèle mécanique. La fonction contrainte u satisfait les équations suivantes:

$$(6.1) \quad \begin{aligned} -\Delta u &= 2 \quad \text{dans } \Sigma, \\ u/\partial\Sigma &= 0, \quad u/\Gamma_0 = c, \\ -\int_{\Gamma_0} \frac{\partial u}{\partial n} d\sigma &= 2\beta_0 \end{aligned}$$

où c est une constante inconnue, $\beta_0 = |\Omega_0|$ l'aire du trou et $|\Sigma| = \beta$ l'aire de la section du matériau.

La rigidité de torsion de P de section Σ a pour expression:

$$(6.2) \quad K(R, \Sigma) = \int_{\Sigma} |\nabla u|^2 dx = 2 \int_{\Sigma} u dx + 2cA_0$$

que nous noterons $K(\Sigma)$ sauf mention du contraire; la fonction $u = u_{\Sigma}$ est la solution de (6.1). Il s'agit de trouver un ouvert Σ tel que

$$(P) \quad K(\Sigma) \geq K(\Sigma')$$

pour tout Σ' ayant les caractéristiques énoncées précédemment. Ce problème entre dans notre cadre d'étude. En effet on plonge Ω_0 dans un "grand ouvert" $\tilde{\Omega} - (|\tilde{\Omega}| \gg A)$ — par exemple une boule $B(O, R) = B$ (Cf. Fig. 3), et on considère le problème de contrôle suivant:

(i) ensemble des contrôles admissibles:

$$(6.3) \quad \mathcal{U} = \left\{ v \in L^{\infty}(\Omega) / 0 \leq v \leq 1, \int_{\Omega} v dx = A \right\}.$$

On prolonge les éléments de \mathcal{U} à Ω_0 par la valeur 1.

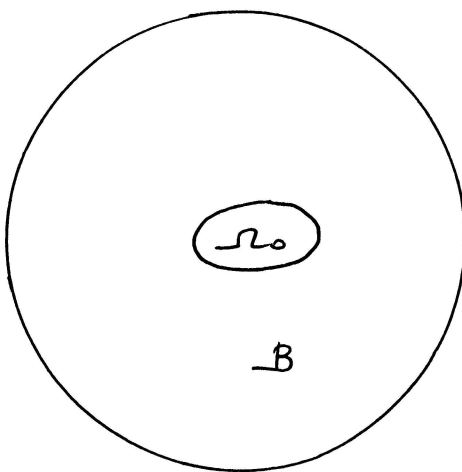


FIG. 3

(ii) La fonction coût:

$$(6.4) \quad J(v) = \int_{\Omega} v \cdot \omega_v dx + 2c_v \cdot A_0$$

où $\omega = \omega_v$ désigne l'état du système, solution de l'équation d'état suivante:

(iii) équation d'état:

$$(6.5) \quad -\Delta \omega = 2v \quad \text{dans } B \setminus \Omega_0 = \Omega$$

$$(6.5) \quad \omega / \Gamma_0 = c, \quad c = \text{constante inconnue}$$

$$\omega / \partial B = 0, \quad \int_{\Gamma_0} \frac{\partial \omega}{\partial \eta} d\sigma = 2 \int_{\Omega_0} v dx = 2\beta_0.$$

Notre problème se formule alors comme suit: trouver $(\bar{v}, \bar{\omega} = \omega_{\bar{v}})$ tel que

$$(6.6) \quad J(\bar{v}) \leq J(v) \quad \forall v \in \mathcal{U}.$$

Remarque. L'unique solution de (6.5) est caractérisée par

$$(6.7) \quad \frac{1}{2} \int_B |\nabla \omega_v|^2 dx - 2 \int_B v \omega_v dx \leq \frac{1}{2} \int_B |\nabla \omega|^2 dx - 2 \int_B v \omega dx$$

pour tout ω dans V où V est l'espace suivant:

$$(6.8) \quad V = \{\omega \in H_0^1(B(0, R)) / \omega = \text{constante sur } \Omega_0\}.$$

Pour l'existence de ce problème on pourra consulter [23].

Pour éviter de se répéter et alourdir notre rédaction, nous ne présenterons pas le détail du résultat d'existence: le lecteur vérifiera que les résultats énoncés dans le cas général sont applicables ici. Il existe une solution optimale $(\bar{v}, \bar{\omega} = \omega_{\bar{v}})$ dont le contrôle \bar{v} est caractérisé par la condition d'optimalité suivante:

$$(6.9) \quad \int_{\Omega} \bar{\omega} \cdot (v - \bar{v}) dx \leq 0 \quad \forall v \in \mathcal{U},$$

condition qui entraîne que \bar{v} possède la propriété bang-bang, propriété qui a été abordée dans un cadre général au § 4. De plus il existe un réel \bar{t} tel que l'ensemble

$$(6.10) \quad \Sigma = \Sigma_{\bar{t}} = \{x \in B(0, R) / \bar{\omega}(x) > \bar{t}\}$$

et le contrôle \bar{v} vérifient:

$$(6.11) \quad \bar{v} = \begin{cases} 1 & \text{sur } \Sigma, \\ 0 & \text{sur } B(0, R) \setminus \bar{\Sigma}; \end{cases}$$

l'optimalité de $(\bar{\omega}, \bar{v})$ entraîne:

$$(6.12) \quad J(\bar{v}) = \frac{1}{2} \int_B |\nabla \bar{\omega}|^2 dx - 2 \int_B \bar{v} \cdot \bar{\omega} \leq \frac{1}{2} \int_B |\nabla \omega|^2 dx - 2 \int_B v \cdot \omega \quad \forall v \in \mathcal{U}, \forall \omega \in V.$$

La question importante qui se pose alors est la suivante: Comment retrouver la formulation initiale de notre problème, autrement dit a-t-on résolu le problème initial?

Nous avons le résultat suivant.

THÉORÈME 6.1. *Si l'on suppose que β est assez petit, alors il existe $\Sigma = \Sigma_R$ ouvert entourant Ω_0 , et une fonction ω_R satisfaisant (6.1) et tels que*

$$K_R = K(\Sigma_R) = \int_{\Sigma_R} |\nabla \omega_R|^2 dx \geq K(\Sigma')$$

pour tout ouvert Σ' satisfaisant $|\Sigma'| = A$, $\Sigma' \subset B(0, R) \setminus \Omega_0$.

Remarque. Nous montrons dans [30] que $c_{\bar{v}} \geq t$; ce qui entraîne que Σ_R entoure Ω_0 . Egalement nous montrons que Σ_R est une couronne simplement connexe.

Démonstration. Il s'agit de construire une fonction état et un contrôle tests utilisables dans le problème de contrôle classique. Considérons les deux problèmes suivants

$$\begin{aligned} -\Delta \omega' &= 2 & \text{dans } \theta = \Omega' \setminus \bar{\Omega}_0, \\ \omega' / \Gamma_0 &= c', & - \int_{\Gamma_0} \frac{\partial \omega'}{\partial \eta} d\sigma = 2\beta_0, \\ \omega' / \partial_1 \theta &= 0; \end{aligned}$$

et

$$\begin{aligned} -\Delta \hat{\omega} &= 0 & \text{dans } B \setminus \bar{\Omega}' \\ \hat{\omega} / \partial_1 \theta &= t', & \hat{\omega} / \partial B = 0 \end{aligned}$$

où t' est un réel qui sera précisé ultérieurement (Cf. Fig. 4).

Et définissons le couple admissible suivant (v_1, ω_1) :

$$\begin{aligned} \omega_1 &= \begin{cases} c' + t' & \text{sur } \Omega_0, \\ \omega' + t' & \text{sur } \theta, \\ \hat{\omega} & \text{sur } B \setminus \bar{\Omega}'; \end{cases} \\ v_1 &= \begin{cases} 1 & \text{sur } \Omega' \\ 0 & \text{sur } B \setminus \bar{\Omega}'. \end{cases} \end{aligned}$$

On utilise ce couple (v_1, w_1) comme fonction test dans (6.12). Après quelques calculs élémentaires on obtient

$$\begin{aligned} (6.13) \quad & \frac{1}{2} \int_{\Sigma} |\nabla(\bar{\omega} - \bar{t})|^2 dx - 2 \int_{\Sigma} (\bar{\omega} - \bar{t}) dx - 2(\bar{c} - \bar{t})\beta_0 - \Delta(t') \\ & \cong \int_{\theta} |\nabla \omega'|^2 dx - 2 \int_{\theta} \omega' dx - 2c'\beta_0 \end{aligned}$$

pour tout $t' > 0$, où

$$\Delta(t') = \frac{1}{2} \int_{B \setminus \bar{\Sigma}} |\nabla \bar{\omega}|^2 - 2\bar{t}(\beta + \beta_0) - \frac{1}{2} \int_{B \setminus \bar{\Omega}'} |\nabla \hat{\omega}|^2 dx + 2t'(\beta + \beta_0).$$

Remarquons que si l'on suppose qu'il existe t' tel que l'on ait

$$(6.14) \quad \Delta(t') \geq 0,$$

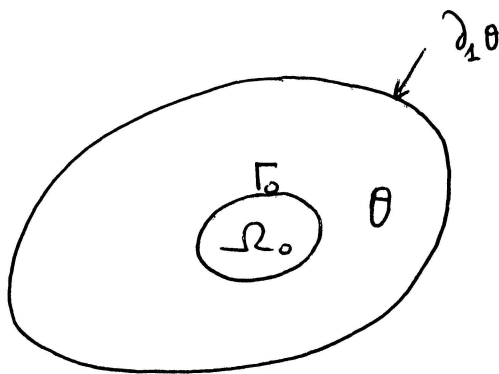


FIG. 4

alors Σ est le domaine optimal cherché, i.e., que (6.13) devient:

$$K(\Sigma) = \int_{\Sigma} |\nabla(\bar{\omega} - \bar{t})|^2 dx \cong \int_{\theta} |\nabla \omega'|^2 dx = K(\Omega')$$

prouvant ainsi que $(\bar{\omega} - \bar{t}, \Sigma)$ est la solution cherchée.

Si l'on suppose, par exemple, que β est assez petit et que les oscillations éventuelles de toute suite de domaines optimisants sont locales, i.e., ne partent pas à l'infini, alors on montre dans [30] que le résultat (6.14) a lieu.

Remarque. Des résultats qualitatifs concernant ce problème seront donnés dans [30].

6.2. Fluide parfait incompressible. On considère un fluide parfait incompressible non soumis à des forces extérieures, occupant un ouvert connexe Ω du plan R^2 (cf. Fig. 5).

Nous nous intéressons aux solutions de l'équation d'Euler stationnaire. Désignons par u la vitesse d'une particule du fluide dans la position (x, y) ; l'incompressibilité se traduit par

$$\operatorname{div} u = 0 \text{ (conservation de la masse).}$$

Il existe des principes variationnels relatifs aux équations d'Euler stationnaires ([27]) (cf. également [29], [31]): les solutions des équations

$$(6.15) \quad \begin{aligned} -u_1 \frac{\partial u_1}{\partial x} - u_2 \frac{\partial u_1}{\partial y} &= \frac{\partial p}{\partial x} \\ -u_1 \frac{\partial u_2}{\partial x} - u_2 \frac{\partial u_2}{\partial y} &= \frac{\partial p}{\partial y} \end{aligned}$$

$$\vec{u} \cdot \vec{n} / \Gamma = 0 \text{ condition de glissement}$$

où p désigne la pression, sont des éléments qui maximisent ou minimisent l'énergie cinétique: [27], [36]

$$E = E(u) = \frac{1}{2} \int_{\Omega} |u|^2 dx.$$

Il faut alors préciser la classe de fonctions sur laquelle on cherche à minimiser ou maximiser E . Pour cela on fait la remarque suivante: ([36], [31]) dans le cas instationnaire entre deux instants t_1 et t_2 la vorticit  du fluide

$$\omega = \operatorname{rot} u$$

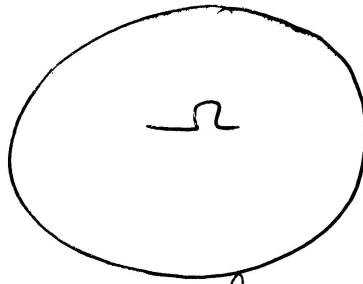


FIG. 5

a un réarrangement fixé. Ce qui définit la nouvelle contrainte:

$$\min \{E(u)/\operatorname{div} u = 0, \text{réarrangement } \omega = v_0: \text{fixé}\}.$$

Transformons quelque peu ce problème: dans \mathbf{R}^2 la condition de conservation de la masse permet d'introduire une fonction courant φ définie à une constante près par

$$\frac{\partial \varphi}{\partial y} = u_1, \quad -\frac{\partial \varphi}{\partial x} = u_2.$$

Nous avons

$$\begin{aligned} \omega &= -\Delta \varphi \\ E &= \frac{1}{2} \int_{\Omega} |\nabla \varphi|^2 dx; \end{aligned}$$

et notre problème s'écrit: il s'agit de maximiser, par exemple, l'énergie

$$\frac{1}{2} \int_{\Omega} |\nabla \varphi|^2 dx$$

sous les contraintes suivantes:

(i)

$$\begin{aligned} -\Delta \varphi &= v \quad (\text{équation d'état}) \\ \varphi/\Gamma &= 0 \end{aligned}$$

(ii)

$$v \in \mathcal{U} = \{v \in L^\infty(\Omega)/\alpha \leq v \leq \beta/v_* = v_0\} \text{ (ensemble des contrôles).}$$

Cet ensemble n'est pas convexe.

Remarque. On peut considérer l'ensemble

$$\mathcal{U} = \{v \in L^p(\Omega)/v_* = v_0\}; \quad p > 1;$$

les résultats démontrés dans les chapitres précédents s'adaptent.

Il existe une solution optimale $(\bar{\varphi}, \bar{v})$ du problème du contrôle ci-dessus; de plus $\bar{v} = f(\bar{\varphi})$ où $f(t) = v_0((\bar{\varphi}^*)^{-1})(t)$ est une fonction décroissante. Vérifions enfin (cela est classique [37]) que \bar{u} défini par

$$\bar{u}_1 = \frac{\partial \bar{\varphi}}{\partial y}, \quad \bar{u}_2 = -\frac{\partial \bar{\varphi}}{\partial x}$$

est solution de (6.15).

Nous avons, formellement

$$\begin{aligned} p_1 &= u_1 \frac{\partial u_1}{\partial x} + u_2 \frac{\partial u_1}{\partial y} \\ &= \frac{\partial \varphi}{\partial y} \frac{\partial^2 \varphi}{\partial x \partial y} - \frac{\partial \varphi}{\partial x} \frac{\partial^2 \varphi}{\partial y^2} \\ &= \frac{\partial \varphi}{\partial y} \frac{\partial^2 \varphi}{\partial x \partial y} + \frac{\partial \varphi}{\partial x} \frac{\partial^2 \varphi}{\partial x^2} - \frac{\partial \varphi}{\partial x} \Delta \varphi \\ &= \frac{1}{2} \frac{\partial}{\partial x} [|\nabla \varphi|^2] + \frac{\partial \varphi}{\partial x} f(\varphi) \\ &= \frac{1}{2} \frac{\partial}{\partial x} [|\nabla \varphi|^2] + \frac{\partial}{\partial x} [F(\varphi)], \end{aligned}$$

où $F(t) = \int_0^t f(s) ds$. De même nous avons

$$\begin{aligned} p_2 &= u_1 \frac{\partial u_2}{\partial x} + u_2 \frac{\partial u_2}{\partial y} \\ &= \frac{1}{2} \frac{\partial}{\partial y} [|\nabla \varphi|^2] + \frac{\partial}{\partial y} [F(\varphi)], \end{aligned}$$

Nous voyons aisément que (p_1, p_2) est un gradient. Et de plus la pression cherchée p a pour expression:

$$-p(x, y) = \frac{1}{2} |\nabla \varphi|^2 + F(\varphi)$$

Remarque. La fonction φ possède la régularité $C^{1,\alpha}(\bar{\Omega}) \cap H^2(\Omega)$, $0 < \alpha < 1$. Ceci permet de justifier les calculs précédents.

Il nous reste à vérifier la condition aux limites

$$u \cdot n = 0 \quad \text{sur } \Gamma.$$

Par (1), pour ψ régulière nous avons $-(u \in (H^1(\Omega))^2) -$:

$$\begin{aligned} \int_{\Omega} [u \cdot \nabla \psi + \psi \operatorname{div} u] dx &= \int_{\Gamma} u \cdot n \psi d\sigma \\ &= \int_{\Omega} \left(\frac{\partial \varphi}{\partial y} \frac{\partial \psi}{\partial x} - \frac{\partial \varphi}{\partial x} \frac{\partial \psi}{\partial y} \right) dx = \int_{\Omega} \left[\frac{\partial}{\partial y} \left(\varphi \frac{\partial \psi}{\partial x} \right) - \frac{\partial}{\partial x} \left(\varphi \frac{\partial \psi}{\partial y} \right) \right] dx \\ &= - \int_{\Gamma} \varphi \left[\frac{\partial \psi}{\partial x} n_2 - \frac{\partial \psi}{\partial y} n_1 \right] d\sigma, \text{ par la formule de Green} \\ &= 0 \text{ puisque } \varphi/\Gamma = 0, \end{aligned}$$

i.e. $\int_{\Gamma} u \cdot n \cdot \psi d\sigma = 0$ pour tout ψ régulière et donc $u \cdot n/\Gamma = 0$ par densité.

Remarque. Dans le cas où le contrôle $v \notin L^\infty(\Omega)$ la régularité précédente n'a pas lieu; mais la preuve de $u \cdot n/\Gamma = 0$ peut encore être justifiée: si $u \in L^2(\Omega)$ est tel que $\operatorname{div} u \in L^2(\Omega)$, on peut encore définir la trace normale $u \cdot n/\Gamma$ [35] (cf. également [29]). \square

Remarque. Des problèmes voisins sont abordés dans [37] et [33] par des méthodes différentes.

ADDITIF:

D'après [12], Auchmuty et Benjamin, motivés par la mécanique des fluides, ont abordé, dans un travail en préparation, la question suivante:

$$\begin{aligned} -\Delta \omega &= v \quad \text{dans } \Omega \\ \omega/\partial\Omega &= 0, \\ \sup \left\{ \int_{\Omega} v \cdot \omega dx / v \cdot t \cdot q \cdot v_* = u_0 \text{ donné} \right\}, \end{aligned}$$

i.e., on prend, dans notre travail du § 5

$$l(v) = v, \quad b(x, \omega) = \omega, \quad c(x, \omega) \equiv 0.$$

Serre [29], avec des motivations voisines de celles de Auchmuty et Benjamin a également abordé un problème assez proche dans le cas où u_0 est une fonction à

valeurs discrètes. Enfin dans [12] on trouve d'autres types de problèmes d'optimisation à réarrangement fixé.

REFERENCES

- [1] I. EKELAND, *Sur le contrôle optimal de systèmes gouvernés par des équations elliptiques*, J. Funct. Anal., 9 (1972), pp. 1-62.
- [2] H. BERLIOCCI AND J. M. LASRY, *Intégrales normales et mesures paramétrées en calcul des variations*, Bull. Soc. Math. France, 101 (1973), pp. 129-184.
- [3] M. F. BIDAUT, *Théorèmes d'existence et d'existence en général d'un contrôle optimal pour des systèmes régis des équations aux dérivées partielles non linéaires*, Thèse, Université de Paris VI, 1973.
- [4] C. CASTAING, *Sur les multiapplications mesurables*, Thèse, Caen, 1967.
- [5] F. MURAT, *Théorèmes de non existence pour des problèmes de contrôle dans les coefficients*, C. R. Acad. Sci. Ser. I Math., 274 (1972), pp. 395-398.
- [6] J. BARANGER, *Existence de solutions pour les problèmes d'optimisation non convexe*, Thèse, Grenoble, 1972.
- [7] L. TARTAR, *Compensated compactness and applications to partial differential equations*, Nonlinear Analysis and Mechanics, Heriot Watt Symposium, Heriot Watt University, Edinburgh, Scotland, UK, 1978, Vol. IV, 136-212.
- [8] J. L. LIONS, *Contrôle optimal*, Dunod, Paris, 1968.
- [9] K. M. CHONG AND N. M. RICE, *Equimeasurable rearrangements of functions*, Queens Papers in Pure and Applied Mathematics, No. 28, Queen's University, Kingston, Ontario, Canada, 1971.
- [10] G. H. HARDY, J. E. LITTLEWOOD, AND G. POLYA, *Inequalities*, Second edition, Cambridge University Press, Cambridge, UK, 1952.
- [11] L. MIGLIACCIO, *Sur une condition de Hardy*, Littlewood, Polya, C. R. Acad. Sci. Ser. I Math., 297 (1983), pp. 25-28.
- [12] A. ALVINO, P. L. LIONS, AND G. TROMBETTI, *On optimization problems with prescribed rearrangements*, Nonlinear Anal. T.M.A. 13 (1989), p. 185-220.
- [13] M. PROTTER AND H. WEINBERGER, *Maximum Principles in Differential Equations*, Springer-Verlag, New York, Berlin, 1984.
- [14] R. TAHRAOUI, *Contrôle optimal à réarrangement fixé*, C. R. Acad. Sci. Paris Ser. I Math, 303 (1986), pp. 955-958.
- [15] ———, *Quelques remarques sur le contrôle des valeurs propres*, Seminaire du Prof. J. L. Lions, College de France, Paris, 1985, Vol. 8, Litman.
- [16] ———, *Contribution à l'étude de quelques questions d'analyse non linéaires issues de la mécanique*, Thèse, Université de Paris VI, 1986.
- [17] M. G. KREIN, *On certain problems of the maximum and the minimum of characteristic values and Lyapunov zones of stability*, Amer. Math. Soc. Transl., 1 (1955), pp. 163-187.
- [18] D. O. BANKS, *Bounds for the eigenvalues of nonhomogeneous hinged vibrating rods*, J. Math. Mech., 16 (1967), pp. 949-966.
- [19] G. ZERAH, *Thèse, Résultats d'existence en forme optimale des poutres*, Université de Paris Sud, 1981.
- [20] R. B. GONZALES DE PAZ, *On the optimal design of elastic shafts*, Math Modelling and Numer. Anal., 23 (1989), pp. 615-625.
- [21] J. CEA, *Problems of shape optimal design*, in Optimization of Distributed Parameter Structures, NATO Proceedings, E. J. Haug and J. Cea, eds., Sijthoff and Noordhoff, Groningen, the Netherlands, 1981.
- [22] D. CHENAIS, *On the existence of a solution in a domain identification problem*, J. Math. Anal. Appl., 52 (1975), pp. 189-219.
- [23] H. LANCHON, *Torsion elastoplastique d'un arbre*, J. Mécan. Theor. Appl., 13 (1974), pp. 267-318.
- [24] O. PIRONNEAU, *Optimal Shape Design for Elliptic Systems*, Springer-Verlag, New York, Berlin, 1984.
- [25] J. P. ZOLESIO, *Identification de domaines par deformation*, Thèse, Université de Nice, 1979.
- [26] N. V. BANICHUK, *Problems and Methods of Optimal Structural Design*, Plenum Press, London, 1983.
- [27] V. I. ARNOLD, *Sur un principe variationnel pour les écoulements stationnaires des liquides parfaits et ses applications aux problèmes de stabilité non linéaire*, J. Mécan. Theor. Appl., 5 (1966), pp. 29-43.
- [28] F. MURAT AND L. TARTAR, *Calcul des variations et homogénéisation*, Ecole d'été d'analyse numérique CEA-EDF-INRIA, pp. 323-367.
- [29] D. SERRE, *Sur la formulation variationnelle de l'écoulement des fluides parfaits*, Equipe D'Analyse Numérique, Lyon-Saint Etienne, 1986.
- [30] R. TAHRAOUI, *Problème à frontière libre et contrôle optimal: résultats qualitatifs*, à paraître.

- [31] T. B. BENJAMIN, *The alliance of practical and analytical insights into the non linear problems of fluid mechanics: Applications of methods of functional analysis to problems in mechanics*, Lecture Notes in Math., Vol. 503, Springer-Verlag, New York, Berlin, 1976, pp. 8–29.
- [32] C. J. AMICK AND L. E. FRAENKEL, *The uniqueness of Hill's spherical vortex*, Arch. Rat. Mech. Anal, 92 (1986), 91–119.
- [33] G. R. BURTON, *Steady symmetric vortex pairs and rearrangements*, Proc. Roy. Soc. Edinburgh Ser. A., 108 (1988) 269–290.
- [34] ———, *Rearrangements of functions, maximisation of convex functional and vortex rings*, Math. Ann., 276 (1987), pp. 225–253.
- [35] R. TEMAM, *Navier-Stokes Equations. Theory and Numerical Analysis*, North-Holland, Amsterdam, 1979.
- [36] V. I. ARNOLD, *Methodes mathématiques de la mécanique classique*, Collection MIR Moscow, 1976.
- [37] B. TURKINGTON, *On steady vortex flow in two dimensions I, II*, Comm. Partial Differential Equations, 8 (1983), pp. 999–1071.
- [38] J. BARANGER AND R. TEMAM, *Nonconvex optimization problems depending on a parameter*, SIAM J. Control., 13 (1970), pp. 146–152.

\mathcal{H}_∞ CONTROLLER SYNTHESIS BY J -LOSSLESS COPRIME FACTORIZATION*

MICHAEL GREEN†

Abstract. This paper develops a coprime factorization approach to the synthesis of internally stabilizing controllers for a given system such that the \mathcal{H}_∞ norm of the closed loop is strictly less than a given bound.

By the use of coprime factorizations, it is shown that the \mathcal{H}_∞ control problem is fundamentally related to the so-called analytic systems considered by Helton et al. [*Regional Conference Series in Mathematics* 68, 1987]. Such problems admit a solution if and only if a certain J -lossless factorization exists. Interpreted in the \mathcal{H}_∞ control context, this means that for a controller of the requisite type to exist the plant must admit a certain J -lossless coprime factorization. The full \mathcal{H}_∞ synthesis problem requires that two nested J -lossless factorizations exist. The results are independent of whether discrete time or continuous time systems are being considered.

It is then shown that J -lossless factorization is equivalent to the existence of a nonnegative, stabilizing solution to an algebraic Riccati equation, allowing a simple state-space formula for a generator of all controllers to be calculated.

Key words. \mathcal{H}_∞ control, coprime factorization, J -spectral factorization, J -contractive matrices, J -lossless matrices

AMS(MOS) subject classifications. 93C35, 93O25, 47B50

1.1. Introduction. This paper concerns the problem of obtaining all stabilizing feedback controllers for a linear time invariant system such that the closed loop transfer function satisfies a prespecified infinity norm bound.

The basic paradigm [8], [11], [12] for solving this problem was to eliminate the internal stability constraint using a doubly coprime factorization and the parametrization of all stabilizing controllers. This leaves a model matching problem which can in turn be reduced to the so-called four block distance problem. Several spectral factorizations enable the four block distance problem to be turned into a Nehari problem, for which solution techniques are known [1], [5], [10], [11], [14]. A variation on this theme involves solving the model matching problem as a bitangential Nevanlinna-Pick interpolation problem [23], [26], [36].

This paradigm is unsatisfactory because the cumbersome chain of reductions, substitutions, factorizations, etc., necessary for the theory is both aesthetically unappealing and leads to computer software that is slow, subject to serious state dimension inflation, and plagued by numerical difficulties, although, in all but the most general case, it was known that controllers of the same degree as the plant exist [26], [29].

Recently the situation has improved dramatically. An elegant solution to the four block distance problem, based on all-pass embedding, has been developed and involves the solution of two algebraic Riccati equations [16], [17], [30]. These equations can also be seen to arise from J -factorizations [16], which have long been associated with the Nehari problem, and form the basis of a solution to the model matching problem [2], [18]. Other new approaches based on the model matching problem have also recently been developed [20], [24].

Separate from this school of thought, consideration was being given to polynomial matrix methods for \mathcal{H}_∞ control [25] and to the attainment of infinity norm objectives by state feedback, where a suitable control is obtained from the solution of a single

* Received by the editors August 22, 1989; accepted for publication (in revised form) December 12, 1990.

† Department of Electrical Engineering, Imperial College, London SW7 2BT, United Kingdom.

algebraic Riccati equation [7], [21], [22], [32], [37]. The full output feedback problem was tackled, although under certain assumptions, in [9], and the solution involved an observer, a state feedback and two Riccati equations.

Connections with other areas of control, such as risk sensitive optimal control [16] and linear quadratic differential games [9], [27], [28], have been established as well as generalizations to time-varying systems [27], [28], [33].

The new observer/state feedback approach of [9] differs fundamentally from the basic paradigm in that it by-passes the parametrization of all stabilizing controllers. Also of interest is that [9] describes, by means of state-space realizations, special problems that are relatively easy to solve, called full information, full control, output estimation, and disturbance feedforward, in terms of which the (restricted) general problem can also be solved. The assumptions made in [9], however, reduce the question of internal stability to input/output stability. So, despite the claim that these assumptions lead to “the simplest special case which captures all the essential features of the general problem,” the role played by internal stability is marginalized.

To fully understand the structure developed in [9], it is essential to find transfer matrix characterizations of the special problems of [9] and to analyze how this impacts upon the internal stability of such systems. This examination leads to the result that the special \mathcal{H}_∞ control problems of [9] are precisely the analytic system problems of Helton et al. [19] in a different form. The connection is revealed by coprime factorization and internal stability. In the general case, the \mathcal{H}_∞ control problem is seen to consist of two nested analytic system problems.

The so-called analytic system problem, for the stable case relevant to \mathcal{H}_∞ control, is the following: Given $G \in \mathcal{H}_\infty$, find all $Q_1, Q_2 \in \mathcal{H}_\infty$ such that

$$\begin{bmatrix} R \\ I \end{bmatrix} = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}, \quad \|R\|_\infty < \gamma.$$

The AS problem can be solved by a number of techniques, ranging from Ball-Helton theory, J -spectral factorization, and interpolation to the state-space approach of [9]. Since the main purpose of this paper is to elucidate the connection via coprime factorization between \mathcal{H}_∞ control and analytic systems, our approach to the AS problem is to show how it relates to model matching and to invoke the results of [18]. This shows that the AS problem has a solution if and only if there exists an outer matrix W such that GW^{-1} is J -lossless. In other words, G has a J -inner outer factorization $G = (GW^{-1})W$.

It follows that the disturbance feedforward \mathcal{H}_∞ problem admits solutions if and only if a right coprime factorization for the plant exists and has a certain J -lossless property. The general \mathcal{H}_∞ synthesis problem is then solved by two nested coprime factorizations, which have J -lossless properties. Since the results are in terms of the properties of transfer matrices, or operators, they apply to both continuous and discrete time systems.

The existence of a J -lossless factorization is shown to be equivalent to the existence of a stabilizing, nonnegative solution to an algebraic Riccati equation with indefinite quadratic term. This allows an explicit state-space formula to be given for a generator of all solutions to the \mathcal{H}_∞ synthesis problem. The state-space results will be proved for the continuous time case and given without detailed proof for discrete time.

1.2. Notation and preliminaries. Vector and matrix notation is standard. $\mathcal{R}^{p \times q}$ denotes the space of proper (bounded at ∞) $p \times q$ rational matrix functions of a complex variable. Let $\mathbb{C}_+ = \{s \in \mathbb{C} : s + \bar{s} \geq 0\} \cup \infty$ and $j\mathbb{R} = \{s \in \mathbb{C} : s + \bar{s} = 0\} \cup \infty$. Let $\mathbb{D}_+ = \{z \in \mathbb{C} : |z| \geq 1\} \cup \infty$ and $\mathbb{T} = \{z \in \mathbb{C} : |z| = 1\}$. Furthermore, let Δ_+ designate either \mathbb{C}_+ or

\mathbb{D}_+ with ∂ denoting either $j\mathbb{R}$ or \mathbb{T} . Denote by $\mathcal{RL}_\infty^{p \times q}(\Delta_+)$ and $\mathcal{RH}_\infty^{p \times q}(\Delta_+)$ the subspaces of $\mathcal{R}^{p \times q}$ without poles in ∂ and Δ_+ , respectively, with $\|\cdot\|_\infty$ the usual associated infinity norm. Denote by $\mathcal{GH}_\infty^p(\Delta_+)$ the group of units of $\mathcal{RH}_\infty^{p \times p}(\Delta_+)$ ($M \in \mathcal{GH}_\infty^p(\Delta_+) \Leftrightarrow M, M^{-1} \in \mathcal{RH}_\infty^{p \times p}$) and let $\gamma\mathcal{BH}_\infty^{p \times q}(\Delta_+) = \{M \in \mathcal{RH}_\infty^{p \times q}(\Delta_+): \|M\|_\infty \leq \gamma\}$. For $\Delta_+ = \mathbb{C}_+$, $\tilde{M}(s) = [M(-\bar{s})]^*$, while, for $\Delta_+ = \mathbb{D}_+$, $\tilde{M}(z) = [M(\bar{z}^{-1})]^*$. $M^*(z) = [M(z)]^*$. To avoid clutter, we will frequently make no explicit reference to Δ_+ or to dimensions of matrices, and speak of \mathcal{R} , \mathcal{RH}_∞ , etc., except where necessary in the interest of clarity or where the result applies only to $\mathcal{RH}_\infty(\mathbb{C}_+)$ or $\mathcal{RH}_\infty(\mathbb{D}_+)$. A matrix $A \in \mathbb{C}^{n \times n}$ is Δ_+ asymptotically stable if none of its eigenvalues lie in Δ_+ .

For $\gamma \neq 0$, define $J_{pq}(\gamma)$ by

$$J_{pq}(\gamma) = \begin{bmatrix} I_p & 0 \\ 0 & -\gamma^2 I_q \end{bmatrix}.$$

We will often abbreviate $J_{pq}(\gamma)$ to J .

A partitioned matrix $M \in \mathcal{R}^{(l+m) \times (q+m)}$ will be called *J-lossless* if $M^* J_{lm}(\gamma) M \leq J_{qm}(\gamma)$ in Δ_+ , with equality holding in ∂ . Also, M is *conjugate J-lossless* if M^* is *J-lossless* (i.e., $MJM^* \leq J$).

As defined, a *J-lossless* matrix need not be square, unlike the situation in [10], [13] (to which several “tricks” in this paper can be traced). Note, however, that if M is *J-lossless*, then $M^* JM$ is invertible on ∂ , so a *J-lossless* matrix is “tall,” i.e., $l \geq q$. Also, note that a *J-lossless* matrix is always assumed to be partitioned, so that the 2,2 corner is square.

Associated with $M \in \mathcal{R}^{p \times q}$ is a state-space realization

$$M(s) = D + C(sI - A)^{-1}B \stackrel{s}{=} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \in \mathbb{C}^{(n+p) \times (n+q)}.$$

Any $M \in \mathcal{R}^{p \times q}$ has right and left coprime factorizations (see [35])

$$M = ND^{-1} = \hat{D}^{-1}\hat{N}, \quad N, D, \hat{N}, \hat{D} \in \mathcal{RH}_\infty, \quad D(\infty), \hat{D}(\infty) \text{ nonsingular}.$$

As in [35], we will abbreviate right coprime as r.c. and right coprime factorization as r.c.f., and similarly for left coprime (l.c.) and left coprime factorization (l.c.f.)

If $P \in \mathcal{R}^{(l+m) \times (p+q)}$ and $K \in \mathcal{R}^{q \times m}$, then the linear fractional map $\mathcal{F}(P, K)$ is defined by

$$\mathcal{F}(P, K) = P_{11} + P_{12}K(I - P_{22}K)^{-1}P_{21}.$$

Note that $\mathcal{F}(P, K)$ is the transfer matrix from w to z in Fig. 1, and that

$$\|\mathcal{F}(P, K)\|_\infty = \sup_{w \neq 0} \frac{\|z\|_2}{\|w\|_2}.$$

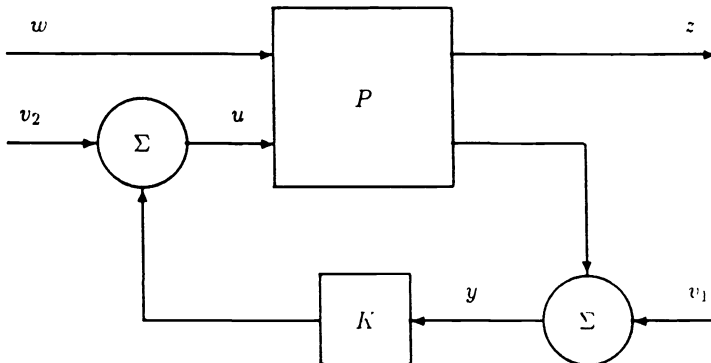


FIG. 1

Thus, $\|\mathcal{F}(P, K)\|_\infty^2$ is the worst case energy gain of the closed loop system. The \mathcal{H}_∞ controller synthesis problem is to establish necessary and sufficient conditions for the existence of stabilizing controllers K satisfying $\|\mathcal{F}(P, K)\|_\infty < \gamma$ and, when these conditions hold, to parametrize all such controllers.

If P_{21} is invertible in \mathcal{R} and G is defined by

$$G = \begin{bmatrix} P_{11} & P_{12} \\ I & 0 \end{bmatrix} \begin{bmatrix} 0 & I \\ P_{21} & P_{22} \end{bmatrix}^{-1} = \begin{bmatrix} I & -P_{11} \\ 0 & P_{21} \end{bmatrix}^{-1} \begin{bmatrix} P_{12} & 0 \\ -P_{22} & I \end{bmatrix},$$

equivalently, G_{22} is invertible in \mathcal{R} and P is defined by

$$(1.1) \quad P = \begin{bmatrix} G_{11} & G_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} G_{21} & G_{22} \\ I & 0 \end{bmatrix}^{-1} = \begin{bmatrix} I & -G_{12} \\ 0 & G_{22} \end{bmatrix}^{-1} \begin{bmatrix} 0 & G_{11} \\ I & -G_{21} \end{bmatrix},$$

then

$$(1.2) \quad \mathcal{F}(P, K) = (G_{11}K + G_{12})(G_{21}K + G_{22})^{-1}.$$

Remark. Equations (1.1) and (1.2) give the basics of the connection between analytic systems, coprime factorization, and \mathcal{H}_∞ control. For if there exists a $G \in \mathcal{RH}_\infty$ such that P has the form (1.1)—which is to say P has a certain coprime factorization over \mathcal{RH}_∞ —we immediately see that

$$\mathcal{F}(P, K) = R_1 R_2^{-1}, \quad \begin{bmatrix} R_1 \\ R_2 \end{bmatrix} = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \begin{bmatrix} \bar{K}_1 \\ \bar{K}_2 \end{bmatrix},$$

where $\bar{K}_1 \bar{K}_2^{-1}$ is a r.c.f. of K . Internal stability amounts to $R_2 \in \mathcal{GH}_\infty$, so we can write the \mathcal{H}_∞ control problem as the analytic system problem

$$\begin{bmatrix} \mathcal{F}(P, K) \\ I \end{bmatrix} = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \begin{bmatrix} K_1 \\ K_2 \end{bmatrix}.$$

2. Stabilization theory. In § 2 we review the stabilization theory of the generalized regulator shown in Fig. 1.

The treatment is based on coprime factorizations over \mathcal{RH}_∞ (see, e.g., [35]).

2.1. The basic stabilization theorem. In this section we give conditions, in terms of coprime factorizations, for a controller to stabilize the generalized regulator. These conditions are given in [11, Chap. 4], but the proof given here is believed to be simpler.

DEFINITION. Let $P \in \mathcal{R}^{(l+m) \times (p+q)}$, $K \in \mathcal{R}^{q \times m}$ and suppose that

$$(2.1) \quad \begin{bmatrix} z \\ y \end{bmatrix} = P \begin{bmatrix} w \\ u \end{bmatrix} + \begin{bmatrix} 0 \\ v_1 \end{bmatrix}, \quad u = Ky + v_2.$$

System (2.1) is well defined if the nine transfer matrices from w, v_1, v_2 to z, y, u exist and are proper (i.e., $\in \mathcal{R}$). The system is *internally stable* if these nine transfer matrices are stable ($\in \mathcal{RH}_\infty$). In this case we say K is a *stabilizing controller* for P , or K *stabilizes* P . We say P is *stabilizable* if there exists a stabilizing controller for P .

THEOREM 2.1. Suppose that $P \in \mathcal{R}^{(l+m) \times (p+q)}$ and let $P = ND^{-1} = \hat{D}^{-1} \hat{N}$ be an r.c.f. and an l.c.f. of P . The following are equivalent:

(i) K stabilizes P ;

$$(2.2) \quad \text{(ii) } K = K_1 K_2^{-1}, \quad \text{with } \begin{bmatrix} D_{11} & D_{12} & 0 \\ D_{21} & D_{22} & K_1 \\ N_{21} & N_{22} & K_2 \end{bmatrix} \in \mathcal{GH}_\infty, K_2(\infty) \text{ nonsingular};$$

$$(2.3) \quad \text{(iii) } K = \hat{K}_2^{-1} \hat{K}_1, \quad \text{with } \begin{bmatrix} \hat{D}_{11} & \hat{D}_{12} & \hat{N}_{12} \\ \hat{D}_{21} & \hat{D}_{22} & \hat{N}_{22} \\ 0 & \hat{K}_1 & \hat{K}_2 \end{bmatrix} \in \mathcal{GH}_\infty, \hat{K}_2(\infty) \text{ nonsingular}.$$

Proof. (i) \Rightarrow (iii). Suppose that K stabilizes P , let K have an l.c.f. $\hat{K}_2^{-1}\hat{K}_1$, and let \hat{X} be the matrix in (2.3). Rewrite (2.1) as

$$(2.4) \quad \hat{X} \begin{bmatrix} z \\ y \\ -u \end{bmatrix} = \hat{Y} \begin{bmatrix} w \\ v_1 \\ -v_2 \end{bmatrix}, \quad \hat{Y} = \begin{bmatrix} \hat{N}_{11} & \hat{D}_{12} & 0 \\ \hat{N}_{21} & \hat{D}_{22} & 0 \\ 0 & 0 & \hat{K}_2 \end{bmatrix}.$$

Since \hat{D} , \hat{N} and \hat{K}_2 , \hat{K}_1 are l.c., it follows that \hat{X} and \hat{Y} are l.c. Thus, by coprimeness, $\hat{X}^{-1}\hat{Y} \in \mathcal{RH}_\infty \Rightarrow \hat{X} \in \mathcal{GH}_\infty$.

(iii) \Rightarrow (i). If (2.3) holds, then it follows from (2.4) that K is a stabilizing controller for P .

(iii) \Rightarrow (ii). Suppose that (2.3) holds. Now let $K = K_1K_2^{-1}$ be an r.c.f. of K . Define

$$X = \begin{bmatrix} -D_{11} & -D_{12} & 0 \\ N_{21} & N_{22} & K_2 \\ D_{21} & D_{22} & K_1 \end{bmatrix}, \quad Y = \begin{bmatrix} -N_{11} & -N_{12} & 0 \\ 0 & 0 & K_2 \\ D_{21} & D_{22} & 0 \end{bmatrix}.$$

Note that X , Y are r.c. and $\hat{X}Y = \hat{Y}X$. Therefore there exist \hat{U} , \hat{V} , U , $V \in \mathcal{RH}_\infty$ such that $\begin{bmatrix} \hat{V} & \hat{U} \\ -\hat{Y} & \hat{X} \end{bmatrix} \begin{bmatrix} X & -U \\ Y & -V \end{bmatrix} = I$. It follows that $X \in \mathcal{GH}_\infty (X^{-1} = \hat{V} + \hat{U}\hat{X}^{-1}\hat{Y})$.

(ii) \Rightarrow (iii). Use a similar argument to that above. \square

2.2. Stabilizability. We now consider necessary conditions on P for a stabilizing controller to exist; we see that P is stabilizable only if P has coprime factorizations in which the denominator matrix has a particular triangular form.

COROLLARY 2.2. *If $P \in \mathcal{R}^{(l+m) \times (p+q)}$ is stabilizable, then*

$$(2.5) \quad (i) \ P \text{ has an r.c.f. } P = ND^{-1}, \text{ with } D = \begin{bmatrix} I_p & 0 \\ D_{21} & D_{22} \end{bmatrix} \text{ and } N_{22}, D_{22} \text{ r.c.};$$

$$(2.6) \quad (ii) \ P \text{ has an l.c.f. } P = \hat{D}^{-1}\hat{N}, \text{ with } \hat{D} = \begin{bmatrix} I_l & \hat{D}_{12} \\ 0 & \hat{D}_{22} \end{bmatrix} \text{ and } \hat{D}_{22}, \hat{N}_{22} \text{ l.c.}$$

Note. Using these factorizations and Theorem 2.1, it is straightforward to show that K stabilizes P if and only if K stabilizes $P_{22} = N_{22}D_{22}^{-1} = \hat{D}_{22}^{-1}\hat{N}_{22}$.

Proof. We prove only that P stabilizable \Rightarrow (i), with (ii) following by similar reasoning.

Suppose that P is stabilizable and let K stabilize P . Let YX^{-1} be an r.c.f. of P , and $K_1K_2^{-1}$ an r.c.f. of K . From Theorem 2.1—see (2.2)—it follows that $[X_{11} \ X_{12}]$ is right invertible in \mathcal{RH}_∞ . Therefore there exist \bar{X}_{21} and \bar{X}_{22} such that $\bar{X} = \begin{bmatrix} X_{11} & X_{12} \\ \bar{X}_{21} & \bar{X}_{22} \end{bmatrix} \in \mathcal{GH}_\infty$. Define $N = Y\bar{X}^{-1}$, $D = X\bar{X}^{-1}$. \square

Remark. The converse to the above corollary is almost true. If P has an r.c.f. $P = ND^{-1}$ as in (2.5), then, since N_{22} , D_{22} are r.c., there exist K_1 , $K_2 \in \mathcal{RH}_\infty$ such that $\begin{bmatrix} D_{22} & K_1 \\ N_{22} & K_2 \end{bmatrix} \in \mathcal{GH}_\infty$. That $K = K_1K_2^{-1}$ stabilizes P follows from Theorem 2.1, provided $K_2(\infty)$ is nonsingular (this is assured if, for example, P_{22} is strictly proper, implying $N_{22}(\infty) = 0$). \square

2.3. Characterization of stabilizing controllers.

COROLLARY 2.3. *Suppose that $P \in \mathcal{R}^{(l+m) \times (p+q)}$ has an r.c.f. $P = ND^{-1}$, $D = \begin{bmatrix} I_p & 0 \\ D_{21} & D_{22} \end{bmatrix}$. Then K stabilizes P if and only if $K = \hat{K}_2^{-1}\hat{K}_1$, where \hat{K}_1 , \hat{K}_2 satisfy*

$$(2.7) \quad \begin{bmatrix} \hat{K}_1 & \hat{K}_2 \end{bmatrix} \begin{bmatrix} N_{21} & -N_{22} \\ -D_{21} & D_{22} \end{bmatrix} = \begin{bmatrix} Q & I_q \end{bmatrix}, \quad \hat{K}_1, \hat{K}_2 \in \mathcal{RH}_\infty, \quad \hat{K}_2(\infty) \text{ nonsingular}.$$

In this case,

$$(2.8) \quad \begin{bmatrix} \mathcal{F}(P, K) \\ I_p \end{bmatrix} = \begin{bmatrix} N_{11} & N_{12} \\ I_p & 0 \end{bmatrix} \begin{bmatrix} I_p \\ Q \end{bmatrix}.$$

Note. The important condition in (2.7) is the second block column; Q is just whatever comes out of the first column once the second is satisfied.

Proof. Suppose that K is a stabilizing controller for P with r.c.f. $K = K_1 K_2^{-1}$. Define

$$\begin{bmatrix} I & 0 & 0 \\ Q & \hat{K}_2 & -\hat{K}_1 \\ \hat{N}_{22} D_{21} - \hat{D}_{22} N_{21} & -\hat{N}_{22} & \hat{D}_{22} \end{bmatrix} = \begin{bmatrix} I & 0 & 0 \\ D_{21} & D_{22} & K_1 \\ N_{21} & N_{22} & K_2 \end{bmatrix}^{-1}.$$

Note that $\hat{K}_2(\infty)$ is nonsingular, since $\hat{K}_2^{-1} = D_{22} - K N_{22}$, and that $K = \hat{K}_2^{-1} \hat{K}_1$. Equation (2.7) follows.

Conversely, suppose that (2.7) holds. Hence \hat{K}_1, \hat{K}_2 are l.c. and N_{22}, D_{22} are r.c., so there exist $\hat{N}_{22}, \hat{D}_{22}, K_1, K_2 \in \mathcal{H}_\infty$ such that

$$\begin{bmatrix} \hat{K}_2 & -\hat{K}_1 \\ -\hat{N}_{22} & \hat{D}_{22} \end{bmatrix} \begin{bmatrix} D_{22} & K_1 \\ N_{22} & K_2 \end{bmatrix} = I.$$

It follows that $K_2(\infty)$ is nonsingular ($K_2^{-1} = \hat{D}_{22} - \hat{N}_{22} \hat{K}_1 \hat{K}_2^{-1}$) and by Theorem 2.1, $K = K_1 K_2^{-1} = \hat{K}_2^{-1} \hat{K}_1$ stabilizes P .

If \hat{K}_2, \hat{K}_1, Q satisfy (2.7), we have

$$\begin{aligned} \mathcal{F}(P, K) &= P_{11} + P_{12}(I - K P_{22})^{-1} K P_{21} \\ &= P_{11} + P_{12} D_{21} + P_{12} D_{22} (\hat{K}_2 D_{22} - \hat{K}_1 N_{22})^{-1} \\ &\quad \times [\hat{K}_1 P_{21} - (\hat{K}_2 D_{22} - \hat{K}_1 N_{22}) D_{22}^{-1} D_{21}] \\ &= P_{11} + P_{12} D_{21} + P_{12} D_{22} [\hat{K}_1 (P_{21} + N_{22} D_{22}^{-1} D_{21}) - \hat{K}_2 D_{21}] \\ &= N_{11} + N_{12} (\hat{K}_1 N_{21} - \hat{K}_2 D_{21}) \\ &= N_{11} + N_{12} Q. \end{aligned}$$

□

Remark. This characterization of all internally stabilizing controllers is implicit, in contrast to the explicit Youla parametrization (see [35]). The advantages for \mathcal{H}_∞ controller synthesis are (1) the parametrization $N_{11} + N_{12} Q$ of the closed loop is “one-sided,” and (2) N_{11} and N_{12} have the same order as P . The Youla parametrization leaves a “two-sided” problem with twice the order of P . Of course, Q in (2.8) is not a free parameter, being constrained by (2.7). The \mathcal{H}_∞ controller synthesis problem is therefore exposed as a two stage procedure, with each stage involving the solution of an analytic systems problem. In stage 1 we find all Q such that $\|N_{11} + N_{12} Q\|_\infty < \gamma$. All such Q 's are given in terms of a free stable matrix U with $\|U\|_\infty < \gamma$. It remains to establish which of these Q 's can be obtained from K via (2.7). Since the map connecting U and Q is invertible, we consider in stage 2 which U 's can be obtained from K by considering the map from K to U , which proves to be of the form in (2.7), that is, another AS problem.

Thus the constraint of internal stability exposes the structure of the general \mathcal{H}_∞ synthesis problem as two nested AS problems—(2.7) a left-handed AS problem, and (2.8) a right-handed AS problem. □

COROLLARY 2.4. Suppose that $P \in \mathcal{R}^{(l+m) \times (p+q)}$ has an l.c.f. $P = \hat{D}^{-1} \hat{N}$, $\hat{D} = \begin{bmatrix} I_l & \hat{D}_{12} \\ 0 & \hat{D}_{22} \end{bmatrix}$. Then K stabilizes P if and only if $K = K_1 K_2^{-1}$, where K_1, K_2 satisfy

$$(2.9) \quad \begin{bmatrix} \hat{Q} \\ I_m \end{bmatrix} = \begin{bmatrix} \hat{N}_{12} & -\hat{D}_{12} \\ -\hat{N}_{22} & \hat{D}_{22} \end{bmatrix} \begin{bmatrix} K_1 \\ K_2 \end{bmatrix} \quad K_1, K_2 \in \mathcal{H}_\infty, \quad K_2(\infty) \text{ nonsingular.}$$

In this case,

$$(2.10) \quad [\mathcal{F}(P, K) \quad I_l] = [I_l \quad \hat{Q}] \begin{bmatrix} \hat{N}_{11} & I_l \\ \hat{N}_{21} & 0 \end{bmatrix} \quad \square$$

2.4. \mathcal{H}_∞ control and analytic systems. Equation (2.10) shows that we can generate the closed loop $\mathcal{F}(P, K)$ from \hat{Q} , with \hat{Q} generated by the system of equations (2.9). Consider now the special case when \hat{N} has the form

$$(2.11) \quad \hat{N} = \begin{bmatrix} 0 & \hat{N}_{12} \\ I_m & \hat{N}_{22} \end{bmatrix}.$$

From (2.10), it follows that $\mathcal{F}(P, K) = \hat{Q}$. We have the following corollary.

COROLLARY 2.5. Suppose that $P \in \mathcal{R}^{(l+m) \times (m+q)}$ has an l.c.f. of the form

$$(2.12) \quad P = \begin{bmatrix} I_l & \hat{D}_{12} \\ 0 & \hat{D}_{22} \end{bmatrix}^{-1} \begin{bmatrix} 0 & \hat{N}_{12} \\ I_m & \hat{N}_{22} \end{bmatrix}.$$

Then K stabilizes P and $\|\mathcal{F}(P, K)\|_\infty \leq \gamma$ if and only if $K = K_1 K_2^{-1}$, where K_1, K_2 satisfy

$$\begin{bmatrix} \hat{Q} \\ I_m \end{bmatrix} = \begin{bmatrix} \hat{N}_{12} & -\hat{D}_{12} \\ -\hat{N}_{22} & \hat{D}_{22} \end{bmatrix} \begin{bmatrix} K_1 \\ K_2 \end{bmatrix} \quad K_1, K_2 \in \mathcal{RH}_\infty, \quad K_2(\infty) \text{ nonsingular}$$

with $\|\hat{Q}\|_\infty \leq \gamma$.

Note. It holds that $\mathcal{F}(P, K) = \hat{Q}$.

Proof. The result follows directly from Corollary 2.4, observing that $\hat{N}_{11} = 0$ and $\hat{N}_{21} = I$. \square

The above result shows that, for certain special systems, characterized by having a coprime factorization of a particular form, there is an exact correspondence between the \mathcal{H}_∞ control problem and the analytic system problem. Naturally, a “left-handed” dual version of Corollary 2.5 is obtained by considering systems with r.c.f. ND^{-1} as in Corollary 2.3 for which $N_{11} = 0$ and $N_{12} = I$.

2.5. Disturbance feedforward and output estimation. Sections 4.3 and 4.4 of [9] concern two special \mathcal{H}_∞ control problems, called disturbance feedforward and output estimation.

For disturbance feedforward, it is assumed that P has a state-space realization

$$(2.13) \quad P \stackrel{s}{=} \begin{bmatrix} A & B_1 & B_2 \\ C_1 & 0 & D_{12} \\ C_2 & I & 0 \end{bmatrix}$$

with $A - B_1 C_2$ asymptotically stable. This gives an l.c.f. as follows:

$$P = \hat{D}^{-1} \hat{N}, \quad \hat{N} \stackrel{s}{=} \begin{bmatrix} A - B_1 C_2 & 0 & B_2 \\ C_1 & 0 & D_{12} \\ C_2 & I & 0 \end{bmatrix}, \quad \hat{D} \stackrel{s}{=} \begin{bmatrix} A - B_1 C_2 & 0 & -B_1 \\ C_1 & I & 0 \\ C_2 & 0 & I \end{bmatrix}.$$

Noting that the states are uncontrollable through the first input, we see that this l.c.f. is of the form in Corollary 2.5.

The output estimation problem is similarly seen to be of the form in Corollary 2.3 with $N_{11} = 0$ and $N_{12} = I$.

The disturbance feedforward problem and the output estimation problem can be completely solved as analytic systems problems. We will see in § 3 that such problems require a single J -spectral factorization, and hence a single Riccati equation for their solution as is already established in [9].

The full information and full control problems of [9] do not appear to arise in a natural way within the framework considered here, since (as is remarked in [9]) they do not satisfy the assumptions that are in force for the general case. Nevertheless, their role in [9] is played by the “model matching” part of the problem—see (2.8) and (2.10). The relationship between disturbance feedforward and full information would seem to be bound up with the reduction of a general analytic system problem to a model-matching type problem. We will see how this is done in § 3.2.

3. Analytic systems. We will now discuss the theory of analytic systems of which those described so far are but a special case. They are stable analytic systems, and the sufficiency theory for such systems can be developed entirely from a consideration of J -lossless matrices. The necessity theory requires a deep theorem, such as the Nehari/AAK, Nevanlinna–Pick, or Ball–Helton theorems, somewhere along the line. To be as concrete as possible, the approach adopted here is to show how the AS problem relates to model matching and to invoke the results of [18]. More sophisticated approaches are possible, and results similar to those presented here can be found in the work of Ball and Helton, especially [4], [19]. Indeed [4] contains many interesting connections between notions of passivity, J -losslessness, analytic systems of various types, the geometry of shift invariant subspaces, and reproducing kernels.

3.1. The analytic system problem. (See [19, p. 57].) Given $G_{11}, G_{12} \in \mathcal{RL}_\infty$, $G_{21}, G_{22} \in \mathcal{RH}_\infty$, an integer $l \geq 0$, and a real number $\gamma > 0$, find the set of all R_1, R_2 such that R_2 is inner and of degree $\leq l$, $\|R_1\|_\infty < \gamma$ and such that there exist $Q_1, Q_2 \in \mathcal{RH}_\infty$ with

$$(3.1) \quad \begin{bmatrix} R_1 \\ R_2 \end{bmatrix} = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}.$$

By suitably choosing the matrix G in (3.1), many problems from operator and analytic function theory—such as Nevanlinna–Pick interpolation, Toeplitz Corona theory and the Nehari model matching and model reduction problems such as are of interest in \mathcal{H}_∞ control—can be posed as AS problems.

Setting $G_{21} = 0$ and $G_{22} = I$, for example, we see that we have the following model matching problem over $\mathcal{RH}_\infty(l)$:

$$(3.2) \quad \begin{bmatrix} R_1 \\ R_2 \end{bmatrix} = \begin{bmatrix} G_{11} & G_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}.$$

Since $R_2 = Q_2$ is inner, it is invertible, and we see that (3.2) is the model matching problem of finding $Q (= Q_1 Q_2^{-1}) \in \mathcal{RH}_\infty(l)$ such that $\|G_{11}Q + G_{12}\|_\infty < \gamma$.

Specializing a little further, setting $G_{11} = I$, we have the Hankel norm model reduction problem, or when $l = 0$, the Nehari problem.

3.2. The case where $l = 0$. With $l = 0$, R_2 is required to be a constant, unitary matrix so, without loss of generality, we will require R_2 to be the identity. Equation (3.1) therefore becomes

$$(3.3) \quad \begin{bmatrix} R \\ I \end{bmatrix} = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}.$$

From the second row of (3.3), it follows that there exist \bar{G}_{11} , \bar{G}_{12} , \bar{Q}_1 , and \bar{Q}_2 such that

$$(3.4) \quad \bar{G} = \begin{bmatrix} \bar{G}_{11} & \bar{G}_{12} \\ G_{21} & G_{22} \end{bmatrix} = \begin{bmatrix} \bar{Q}_1 & Q_1 \\ \bar{Q}_2 & Q_2 \end{bmatrix}^{-1} \in \mathcal{GH}_\infty.$$

This gives

$$(3.5) \quad G = G\bar{G}^{-1}\bar{G} = \begin{bmatrix} \hat{G}_{11} & \hat{G}_{12} \\ 0 & I \end{bmatrix} \bar{G},$$

and, since $\bar{G} \in \mathcal{GH}_\infty$, the analytic system (3.3) is equivalent to the analytic system

$$(3.6) \quad \begin{bmatrix} R \\ I \end{bmatrix} = \begin{bmatrix} \hat{G}_{11} & \hat{G}_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} \hat{Q}_1 \\ I \end{bmatrix} \quad \text{with} \quad \begin{bmatrix} \hat{Q}_1 \\ I \end{bmatrix} = \bar{G} \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}.$$

Thus all $l=0$ AS problems are model matching problems.

THEOREM 3.1 (See [18]). *Suppose that $\hat{G} = \begin{bmatrix} \hat{G}_{11} & \hat{G}_{12} \\ 0 & I_m \end{bmatrix} \in \mathcal{RL}_\infty^{(1+m) \times (q+m)}$ is left invertible in \mathcal{RL}_∞ . There exists a $\hat{Q} \in \mathcal{RH}_\infty$ such that*

$$(3.7) \quad \|\hat{G}_{11}\hat{Q} + \hat{G}_{12}\|_\infty < \gamma$$

if and only if there exists a \hat{W} such that

$$(3.8) \quad \hat{G}^* J_{lm}(\gamma) \hat{G} = \hat{W}^* J_{qm}(\gamma) \hat{W}, \quad \hat{W} \in \mathcal{GH}_\infty^{q+m}, \quad \text{and} \quad \hat{W}_{11} \in \mathcal{GH}_\infty^q.$$

In this case, all $\hat{Q} \in \mathcal{RH}_\infty$ satisfying $\|\hat{G}_{11}\hat{Q} + \hat{G}_{12}\|_\infty \leq \gamma$ are generated by

$$(3.9) \quad \hat{Q} = \hat{Q}_1 \hat{Q}_2^{-1}, \quad \begin{bmatrix} \hat{Q}_1 \\ \hat{Q}_2 \end{bmatrix} = \hat{W}^{-1} \begin{bmatrix} U \\ I_m \end{bmatrix}, \quad U \in \gamma \mathcal{BH}_\infty^{q \times m},$$

where \hat{W} is any solution to (3.8).

Furthermore, for any \hat{W} satisfying (3.8), the following hold:

- (a) *For all $U \in \gamma \mathcal{BH}_\infty^{q \times m}$, \hat{Q}_2 in (3.9) $\in \mathcal{GH}_\infty^m$;*
- (b) *For all $\hat{Q} \in \mathcal{RH}_\infty^{q \times m}$ satisfying $\|\hat{G}_{11}\hat{Q} + \hat{G}_{12}\|_\infty \leq \gamma$, $[0 \ I_m] \hat{W} \begin{bmatrix} \hat{Q} \\ I_m \end{bmatrix} \in \mathcal{GH}_\infty^m$.*

Proof. Variations of this result appear in a number of places; see, for example, [2], [3], [4], [19], or [18], from which the continuous time ($\Delta_+ = \mathbb{C}_+$) version of the above statement derives. The discrete time result follows by bilinear transformation. The Nehari case is treated in [5], [10], [11]. \square

COROLLARY 3.2. *Suppose that $G \in \mathcal{RL}_\infty^{(1+m) \times (q+m)}$ is left invertible in \mathcal{RL}_∞ , $[0 \ I_m]G \in \mathcal{RH}_\infty^{m \times (q+m)}$. There exist $Q_1, Q_2 \in \mathcal{RH}_\infty$ and an R with $\|R\|_\infty < \gamma$ satisfying the analytic system (3.3) if and only if there exists a W such that*

$$(3.10) \quad \hat{G}^* J_{lm}(\gamma) G = W^* J_{qm}(\gamma) W, \quad W \in \mathcal{GH}_\infty^{q+m}, \quad \text{and} \quad (GW^{-1})_{22} \in \mathcal{GH}_\infty^m.$$

In this case, all solutions R to the analytic system (3.3) are generated by

$$(3.11) \quad R = R_1 R_2^{-1}, \quad \begin{bmatrix} R_1 \\ R_2 \end{bmatrix} = GW^{-1} \begin{bmatrix} U \\ I_m \end{bmatrix}, \quad U \in \gamma \mathcal{BH}_\infty^m$$

with W any solution to (3.10).

Proof. First note that if a solution to the AS exists or a W satisfying (3.10) exists, then $[G_{21} \ G_{22}]$ is right invertible in \mathcal{RH}_∞ . The result follows from (3.4)-(3.6) and Theorem 3.1. Note that $\hat{W}_{11} \in \mathcal{GH}_\infty \Leftrightarrow (\hat{W}^{-1})_{22} \in \mathcal{GH}_\infty$ and that $(GW^{-1})_{22} = (\hat{W}^{-1})_{22}$. \square

3.3. J -lossless matrices and stable analytic systems. If $G \in \mathcal{RH}_\infty$, the conditions of Corollary 3.2 can be given an attractive, alternative form using J -lossless matrices via the following lemma.

LEMMA 3.3. Suppose that $X \in \mathcal{RH}_\infty^{(l+m) \times (q+m)}$. Then $X_{22} \in \mathcal{GH}_\infty^m$ and $X^* J_{lm}(\gamma) X = J_{qm(\gamma)}$ if and only if X is J -lossless.

Proof. Suppose that X is J -lossless. Then $X_{12}^* X_{12} - \gamma^2 X_{22}^* X_{22} \leq -\gamma^2 I$ in Δ_+ , so X_{22} is nonsingular in Δ_+ , i.e., $X_{22} \in \mathcal{GH}_\infty$. By J -losslessness, we have that $X^* JX = J$ on ∂ . Therefore, since X is rational, so $X^* JX$ is rational, we have that $X^* JX = J$.

Conversely, suppose that $X_{22} \in \mathcal{GH}_\infty$ and $X^* JX = J$. Let

$$X_1 = \begin{bmatrix} X_{11} & X_{12} \\ 0 & \gamma I \end{bmatrix}, \quad X_2 = \begin{bmatrix} I & 0 \\ \gamma X_{21} & \gamma X_{22} \end{bmatrix}.$$

Note that $X_2 \in \mathcal{GH}_\infty$ and $Y = X_1 X_2^{-1} \in \mathcal{RH}_\infty$. Now observe that

$$(3.12) \quad X' JX - J = X_2' \{ Y' Y - I \} X_2,$$

where X' denotes either X^* or X^* . With X' denoting X^* , we see that Y is all-pass. Since Y is stable, we have, by the maximum modulus principle, that Y is lossless ($Y^* Y \leq I$ in Δ_+). Therefore, from (3.12) with X' denoting X^* , X is J -lossless. \square

Before interpreting Theorem 3.2 in terms of J -lossless factorizations, we have the following generalization of Lemma 3.3, which is proved using Theorem 3.1. It can be proved independently, giving a sufficiently theory for stable analytic systems, but since we also need necessity, it is more efficient to use previously established results.

LEMMA 3.4. Let $X \in \mathcal{RH}_\infty^{(l+m) \times (q+m)}$ be J -lossless and define G and W by

$$W = \begin{bmatrix} I_q & 0 \\ X_{21} & X_{22} \end{bmatrix}^{-1}, \quad G = XW = \begin{bmatrix} G_{11} & G_{12} \\ 0 & I_m \end{bmatrix}.$$

(1) If $U \in \gamma \mathcal{BH}_\infty$, then $X_{21}U + X_{22} \in \mathcal{GH}_\infty$ and $(X_{11}U + X_{12})(X_{21}U + X_{22})^{-1} \in \gamma \mathcal{BH}_\infty$.

(2) If $Q \in \mathcal{RH}_\infty$ is such that $\|G_{11}Q + G_{12}\|_\infty \leq \gamma$, then $I - X_{21}Q \in \mathcal{GH}_\infty$ and $Q(I - X_{21}Q)^{-1}X_{22} = (I - QX_{21})^{-1}QX_{22} \in \gamma \mathcal{BH}_\infty$.

Proof. Since X is J -lossless, $G^* JG = W^* JW$ and $W \in \mathcal{GH}_\infty$, by Lemma 3.3. Clearly, $W_{11} \in \mathcal{GH}_\infty$. The result follows from Theorem 3.1—see (a) and (b). \square

THEOREM 3.5. Suppose that $G \in \mathcal{RH}_\infty^{(l+m) \times (q+m)}$ is left invertible in \mathcal{RL}_∞ . Then there exist $Q_1, Q_2 \in \mathcal{RH}_\infty$ and an R with $\|R\|_\infty < \gamma$ satisfying the analytic system (3.3) if and only if there exists a W such that

$$(3.13) \quad W \in \mathcal{GH}_\infty^{q+m} \text{ and } GW^{-1} \text{ is } J\text{-lossless}^1.$$

In this case, all solutions R to the analytic system (3.3) with $\|R\|_\infty \leq \gamma$ are generated by

$$R = R_1 R_2^{-1}, \quad \begin{bmatrix} R_1 \\ R_2 \end{bmatrix} = GW^{-1} \begin{bmatrix} U \\ I_m \end{bmatrix}, \quad U \in \gamma \mathcal{BH}_\infty^{q+m},$$

where W is any solution to (3.13). All Q_1, Q_2 are generated by

$$\begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} R_2 = W^{-1} \begin{bmatrix} U \\ I_m \end{bmatrix}, \quad U \in \gamma \mathcal{BH}_\infty^{q+m}.$$

¹ The theorem is stated with G assumed to be partitioned so that the 2,2 corner is square. We also need to consider the case when the 2,1 is square, i.e., $G \in \mathcal{RH}_\infty^{(l+m) \times (m+q)}$. One is converted to the other by swapping columns, and the off-diagonal blocks of W are then square, instead of the diagonal blocks.

Proof. The proof follows from Corollary 3.2 and Lemma 3.3. Note also that by Lemma 3.4, $R_2 \in \mathcal{GH}_\infty$. \square

3.4. Some comments concerning infinity. In the application of the theory of analytic systems to \mathcal{H}_∞ control, $K = K_1 K_2^{-1}$ is a stabilizing controller for a problem of disturbance feedforward type if and only if $\begin{bmatrix} K_1 \\ K_2 \end{bmatrix}$ satisfy an analytic system and $K_2(\infty)$ is nonsingular, or else K is not proper. That is, infinity is a special point because we allow the controller to have poles anywhere in \mathbb{C} but not at infinity.

The following lemma allows us to say that if $Q_2(\infty)$ is nonsingular in Theorem 3.5 for some strict contraction $U(\infty)$,² then we can assume $(W^{-1})_{22}(\infty)$ is nonsingular. This will mean that $Q_2(\infty)$ will be nonsingular for any U with $U(\infty) = 0$.

LEMMA 3.6. *Let $\bar{\Omega} \in \mathbb{C}^{(q+m) \times (q+m)}$ be nonsingular. There exists a constant matrix $U \in \mathbb{C}^{q \times m}$ such that $U^* U < \gamma^2 I$ and $\begin{bmatrix} 0 & I_m \end{bmatrix} \bar{\Omega} \begin{bmatrix} U \\ I_m \end{bmatrix}$ is nonsingular if and only if there exists a nonsingular matrix $\Omega \in \mathbb{C}^{(q+m) \times (q+m)}$ such that $\Omega J^{-1} \Omega^* = \bar{\Omega} J^{-1} \bar{\Omega}^*$, with $\Omega_{22} \in \mathbb{C}^{q \times q}$ nonsingular.*

Proof. If Ω exists, then $\Omega = \bar{\Omega} \Theta$, where $\Theta J^{-1} \Theta^* = J^{-1}$. This implies $\Theta_{12}^* \Theta_{12} - \gamma^{-2} \Theta_{22}^* \Theta_{22} = -\gamma^{-2} I$, so Θ_{22} is nonsingular and $U = \gamma^2 \Theta_{12} \Theta_{22}^{-1}$ satisfies $U^* U < \gamma^2 I$. We then have $\begin{bmatrix} 0 & I_m \end{bmatrix} \bar{\Omega} \begin{bmatrix} U \\ I_m \end{bmatrix} = \gamma^2 \Omega_{22} \Theta_{22}^{-1}$, which is nonsingular.

Conversely, if U exists, set $\Omega = \bar{\Omega} \mathcal{J}(U)$, where

$$\mathcal{J}(U) = \begin{bmatrix} (\gamma^2 I - UU^*)^{-1/2} & 0 \\ 0 & (\gamma^2 I - U^* U)^{-1/2} \end{bmatrix} \begin{bmatrix} \gamma I & \gamma U \\ \gamma^{-1} U^* & \gamma \end{bmatrix}.$$

We then have $\mathcal{J}(U) J^{-1} \mathcal{J}(U)^* = J^{-1}$ and that $\begin{bmatrix} 0 & I \end{bmatrix} \Omega^{-1} \begin{bmatrix} 0 \\ I \end{bmatrix} = \begin{bmatrix} 0 & I \end{bmatrix} \bar{\Omega}^{-1} \begin{bmatrix} U \\ I \end{bmatrix} \gamma (\gamma^2 I - U^* U)^{-1/2}$. \square

A further result we need for similar reasons is the following lemma.

LEMMA 3.7. *Suppose that $G \in \mathcal{RH}_\infty^{(l+m) \times (q+m)}$ with G_{21} strictly proper³ (or $G \in \mathcal{RH}^{(l+m) \times (m+q)}$ with G_{22} strictly proper) and there exists a \hat{W} satisfying (3.13). Then there exists a W satisfying (3.13) with W_{21} (respectively, W_{22}) strictly proper.*

Note. $W(\infty)$ nonsingular and $W_{21}(\infty) = 0$ (respectively, $W_{22}(\infty) = 0$) implies that $W_{11}(\infty)$ and $W_{22}(\infty)$ (respectively, $W_{21}(\infty)$ and $W_{12}(\infty)$) are nonsingular.

Proof. Consider the $G_{21}(\infty) = 0$ case.

For $\Delta_+ = \mathbb{C}_+$, since $\infty \in \partial$, we have $[\hat{W}^* J \hat{W}](\infty) = [G^* J G](\infty)$. Since $\hat{W}(\infty)$ is nonsingular, $[G^* J G](\infty)$ is nonsingular and it follows that $\{[G^* J G](\infty)\}_{11} > 0$. Now by taking the Schur complement, $W(\infty)$ with the desired $W_{21}(\infty) = 0$ property exists. Define $W = W(\infty) \hat{W}(\infty)^{-1} \hat{W}$.

For $\Delta_+ = \mathbb{D}_+$, since $\infty \in \mathbb{D}_+$, we have $[\hat{W}^* J \hat{W}](\infty) \geq [G^* J G](\infty)$. Let Λ be such that $[\hat{W}^* J \hat{W}](\infty) = [G^* J G](\infty) + \Lambda^* \Lambda = [\hat{G}^* \hat{J} \hat{G}](\infty)$, where $\hat{G}^*(\infty) = [\Lambda^* G^*(\infty)]$ and $\hat{J} = I \oplus J$. Now apply the same construction as for the \mathbb{C}_+ case. \square

3.5. Solution to \mathcal{H}_∞ control problems of analytic system type. Theorem 3.5 can now be used to solve the special \mathcal{H}_∞ control problems, which are of stable analytic system type described in § 2.4.

For disturbance feedforward problems, we have the following theorem.

THEOREM 3.8. *Let $P \in \mathcal{R}^{(l+m) \times (m+q)}$ have l.c.f.*

$$P = \begin{bmatrix} I_l & \hat{D}_{12} \\ 0 & \hat{D}_{22} \end{bmatrix}^{-1} \begin{bmatrix} 0 & \hat{N}_{12} \\ I_m & \hat{N}_{22} \end{bmatrix} \quad \text{with} \quad G = \begin{bmatrix} \hat{N}_{12} & -\hat{D}_{12} \\ -\hat{N}_{22} & \hat{D}_{22} \end{bmatrix}$$

left invertible in \mathcal{RL}_∞ .

² Since $\infty \in \Delta_+$, $U(\infty)$ is a strict contraction provided U is a strict contraction on Δ_+ .

³ G_{21} strictly proper means $G_{21}(\infty) = 0$.

There exists an internally stabilizing controller K such that $\|\mathcal{F}(P, K)\|_\infty < \gamma$ if and only if P has an r.c.f.

$$(3.14) \quad P = ND^{-1}, \quad \text{with} \quad \begin{bmatrix} N_{11} & N_{12} \\ D_{11} & D_{12} \end{bmatrix} J\text{-lossless and } N_{22}(\infty) \text{ nonsingular.}^4$$

In this case, all stabilizing controllers K such that $\|\mathcal{F}(P, K)\|_\infty \leq \gamma$ are given by $K = K_1 K_2^{-1}$, where K_1, K_2 are such

$$(3.15) \quad \begin{bmatrix} K_1 \\ K_2 \end{bmatrix} = \begin{bmatrix} D_{21} & D_{22} \\ N_{21} & N_{22} \end{bmatrix} \begin{bmatrix} U \\ I_m \end{bmatrix}, \quad U \in \gamma \mathcal{B}\mathcal{H}_\infty^{q \times m}, \quad K_2(\infty) \text{ nonsingular.}$$

Note. Since $N_{22}(\infty)$ is nonsingular, $K_2(\infty)$ in (3.15) is nonsingular for all strictly proper U ($U(\infty) = 0$).

Proof. First observe the following relationships: Rewrite the l.c.f. $\hat{D}^{-1}\hat{N}$ as the r.c.f. YX^{-1}

$$P = YX^{-1}, \quad Y = \begin{bmatrix} \hat{N}_{12} & -\hat{D}_{12} \\ 0 & I \end{bmatrix}, \quad X = \begin{bmatrix} -\hat{N}_{22} & \hat{D}_{22} \\ I & 0 \end{bmatrix}.$$

Now

$$P = ND^{-1} \text{ is a r.c.f. with } N_{22}(\infty) \text{ nonsingular}$$

$$\Leftrightarrow \exists W \in \mathcal{GH}_\infty \text{ with } YW^{-1} = N \text{ and } XW^{-1} = D, \quad (W^{-1})_{22}(\infty) \text{ nonsingular.}$$

Observing the structure of Y and X , we see that such a W exists if and only if

$$(3.16) \quad W^{-1} = \begin{bmatrix} D_{21} & D_{22} \\ N_{21} & N_{22} \end{bmatrix} \in \mathcal{GH}_\infty, \quad N_{22}(\infty) \text{ nonsingular}$$

and

$$(3.17) \quad G = \begin{bmatrix} \hat{N}_{12} & -\hat{D}_{12} \\ -\hat{N}_{22} & \hat{D}_{22} \end{bmatrix} = \begin{bmatrix} N_{11} & N_{12} \\ D_{11} & D_{12} \end{bmatrix} W.$$

Obviously, GW^{-1} J -lossless $\Leftrightarrow \begin{bmatrix} N_{11} & N_{12} \\ D_{11} & D_{12} \end{bmatrix}$ J -lossless.

Suppose that there exists a stabilizing controller K for P with $\|\mathcal{F}(P, K)\|_\infty < \gamma$. Then, by Corollary 2.5 and Theorem 3.5, there exists a $W \in \mathcal{GH}_\infty$ such that GW^{-1} is J -lossless, with G as in the first equality in (3.17). Also $K = K_1 K_2^{-1}$, with K_1, K_2 given by

$$\begin{bmatrix} K_1 \\ K_2 \end{bmatrix} = W^{-1} \begin{bmatrix} U \\ I \end{bmatrix}, \quad \text{some } U \in \mathcal{RH}_\infty, \quad \|U\|_\infty < \gamma.$$

Since K is proper, $K_2(\infty)$ is nonsingular. By Lemma 3.6 we can therefore assume $(W^{-1})_{22}(\infty)$ is nonsingular. Define $N = YW^{-1}$, $D = DW^{-1}$.

Conversely, suppose that P has an r.c.f. as in (3.14). Define W and G as in (3.16), (3.17). By Theorem 3.5 the analytic system defined by G has a solution, and we obtain a suitable controller for P by picking a U in (3.17) with $\|U\|_\infty < \gamma$. Note that, since $N_{22}(\infty)$ is nonsingular, $K_2(\infty)$ is nonsingular for any U with $U(\infty) = 0$.

By Corollary 2.5 and Theorem 3.5, all solutions are generated by (3.15). \square

⁴ $N \in \mathcal{RH}_\infty^{(l+m) \times (q+m)}$, $D \in \mathcal{RH}_\infty^{(m+q) \times (q+m)}$. That is, D_{12} and D_{21} are square, instead of the more usual situation where D_{11} and D_{22} are square. Some ‘‘swapping’’ inevitably occurs; we either do it here, or in Corollaries 2.3, 2.4, 2.5, etc. One reason for doing it this way is the more natural connection with the state-space notation.

For output estimation problems, we have the following theorem.

THEOREM 3.9. *Let $P \in \mathcal{R}^{(q+m) \times (p+q)}$ have r.c.f.*

$$P = \begin{bmatrix} 0 & I_q \\ N_{21} & N_{22} \end{bmatrix} \begin{bmatrix} I_p & 0 \\ D_{21} & D_{22} \end{bmatrix}^{-1}$$

with $G = \begin{bmatrix} N_{21} & -N_{22} \\ -D_{21} & D_{22} \end{bmatrix}$ right invertible in \mathcal{RL}_∞ .

There exists a stabilizing controller K such that $\|\mathcal{F}(P, K)\|_\infty < \gamma$ if and only if P has an l.c.f.

$$(3.18) \quad P = \hat{D}^{-1} \hat{N}, \quad \text{with} \quad \begin{bmatrix} \hat{N}_{11} & \hat{D}_{11} \\ \hat{N}_{21} & \hat{D}_{21} \end{bmatrix} \text{ conjugate } J\text{-lossless and } \hat{N}_{22}(\infty) \text{ nonsingular.}^5$$

In this case, all internally stabilizing controllers such that $\|\mathcal{F}(P, K)\|_\infty \leq \gamma$ are given by

$$(3.19) \quad K = \hat{K}_2^{-1} \hat{K}_1,$$

where \hat{K}_2, \hat{K}_1 are such that

$$(3.20) \quad [\hat{K}_1 \quad \hat{K}_2] = [U \quad I_q] \begin{bmatrix} \hat{D}_{12} & \hat{N}_{12} \\ \hat{D}_{22} & \hat{N}_{22} \end{bmatrix}, \quad U \in \gamma \mathcal{BH}_\infty^{q \times m}, \quad \hat{K}_2(\infty) \text{ nonsingular.}$$

4. \mathcal{H}_∞ control. Although certain \mathcal{H}_∞ control problems, as discussed in § 3.5, can be solved directly as analytic system problems, not all problems are of this type. In this section we treat the general case by observing that a necessary condition for a solution to exist can be used to turn the general problem into an equivalent AS type problem (of output estimation type), which can be solved using Theorem 3.9.

4.1. A necessary condition. By hypothesizing the existence of a stabilizing controller K such that $\|\mathcal{F}(P, K)\|_\infty < \gamma$, we see that a certain analytic system of the model matching type must have a solution. This implies that a certain J -lossless factorization exists. This is used to show that a certain right coprime factorization of P exists.

LEMMA 4.1. *Suppose that $P \in \mathcal{R}^{(l+m) \times (p+q)}$ has an r.c.f. $P = YX^{-1}$ such that $X = \begin{bmatrix} I_p & 0 \\ X_{21} & X_{22} \end{bmatrix}$, Y_{22}, X_{22} are r.c., and Y_{12} is left invertible in \mathcal{RL}_∞ .*

If there exists a stabilizing controller K for P such that $\|\mathcal{F}(P, K)\|_\infty < \gamma$, then P has an r.c.f.

$$(4.1) \quad P = ND^{-1}, \quad \text{with} \quad \begin{bmatrix} N_{11} & N_{12} \\ D_{11} & D_{12} \end{bmatrix} J\text{-lossless and } D_{21}(\infty) \text{ nonsingular.}^6$$

Proof. By Corollary 2.3 there exists a $Q \in \mathcal{RH}_\infty$ such that

$$\begin{bmatrix} \mathcal{F}(P, K) \\ I \end{bmatrix} = \begin{bmatrix} Y_{11} & Y_{12} \\ I & 0 \end{bmatrix} \begin{bmatrix} I \\ Q \end{bmatrix} = G \begin{bmatrix} I \\ Q \end{bmatrix}.$$

By Theorem 3.5 there exists a $W \in \mathcal{GH}_\infty$ such that GW^{-1} is J -lossless. Define $N = YW^{-1}$, $D = XW^{-1}$. By Lemma 3.7 we can take $W_{22}(\infty) = 0$, and it follows that $D_{21}(\infty) = X_{22}(\infty)W_{12}(\infty)^{-1}$. \square

Remarks. If a stabilizing controller exists, then P has a factorization YX^{-1} , with X lower triangular as in the lemma. The condition that Y_{12} have a left inverse in \mathcal{RL}_∞ is, however, an additional assumption.

It is easy to see that if $YX^{-1} = \hat{Y}\hat{X}^{-1}$ are two lower triangular r.c. factorizations, then Y_{12} has a left inverse in \mathcal{RL}_∞ if and only if \hat{Y}_{12} has a left inverse in \mathcal{RL}_∞ . ($\hat{Y}_{12} = Y_{12}X_{22}^{-1}\hat{X}_{22}$. Since $Y_{22}X_{22}^{-1} = \hat{Y}_{22}\hat{X}_{22}^{-1}$ are r.c.f.'s of P_{22} , $X_{22}^{-1}\hat{X}_{22} \in \mathcal{GH}_\infty$.)

⁵ $\hat{N} \in \mathcal{RH}_\infty^{(m+q) \times (p+q)}$, $\hat{D} \in \mathcal{RH}_\infty^{(m+q) \times (q+m)}$.

⁶ $N \in \mathcal{R}^{(l+m) \times (q+p)}$, $D \in \mathcal{R}^{(p+q) \times (q+p)}$, so the off-diagonal blocks of D are square.

Note also that, for Y_{12} to have a left inverse in \mathcal{RL}_∞ , it is necessary, but not sufficient, that P_{12} have a left inverse in \mathcal{RL}_∞ . (It is sufficient if $[Y_{12}]_{X_{22}}^{Y_{12}}$ has full column rank on ∂ : $P_{12} = Y_{12}X_{22}^{-1}$, so for $\lambda \in \partial$, $Y_{12}(\lambda)x = 0 \Rightarrow X_{22}(\lambda)x = y \neq 0$, so $P_{12}(\lambda)y = Y_{12}(\lambda)x = 0$.)

Remark. The best control, the worst disturbance, and the equilibrium equation of [9], [27] are easily seen from the factorization $P = ND^{-1}$ as follows:

$$\begin{bmatrix} u - u_{\text{opt}} \\ w - w_{\text{worst}} \end{bmatrix} = D^{-1} \begin{bmatrix} w \\ u \end{bmatrix} \Rightarrow \begin{bmatrix} z \\ w \end{bmatrix} = \begin{bmatrix} N_{11} & N_{12} \\ D_{11} & D_{12} \end{bmatrix} \begin{bmatrix} u - u_{\text{opt}} \\ w - w_{\text{worst}} \end{bmatrix}.$$

By J -losslessness, we have

$$\|z\|_2^2 - \gamma^2 \|w\|_2^2 = \|u - u_{\text{opt}}\|_2^2 - \gamma^2 \|w - w_{\text{worst}}\|_2^2.$$

The control u_{opt} is optimal in the entropy sense [15]. To obtain an explicit formula for u_{opt} , set $u = u_{\text{opt}}$ above:

$$\begin{bmatrix} w \\ u_{\text{opt}} \end{bmatrix} = D \begin{bmatrix} 0 \\ w - w_{\text{worst}} \end{bmatrix}$$

Thus

$$u_{\text{opt}} = D_{22}(w - w_{\text{worst}}) = D_{22}D_{12}^{-1}w = -(D^{-1})_{12}^{-1}(D^{-1})_{11}w.$$

Note. D_{12} is nonsingular, since it is the 2,2 block of a J -lossless matrix.

By using a left coprime factorization of P , we have the dual of Lemma 4.1, below.

LEMMA 4.2. Suppose that $P \in \mathcal{R}^{(l+m) \times (p+q)}$ has an l.c.f. $P = \hat{X}^{-1}\hat{Y}$ such that $\hat{X} = \begin{bmatrix} I_l & \hat{X}_{12} \\ 0 & \hat{X}_{22} \end{bmatrix}$, \hat{Y}_{22} , \hat{X}_{22} are l.c., and \hat{Y}_{21} is right invertible in \mathcal{RL}_∞ .

If there exists a stabilizing controller K such that $\|\mathcal{F}(P, K)\|_\infty < \gamma$, then P has an l.c.f.

$$(4.2) \quad P = \hat{D}^{-1}\hat{N}, \quad \text{with} \quad \begin{bmatrix} \hat{N}_{11} & \hat{D}_{11} \\ \hat{N}_{21} & \hat{D}_{21} \end{bmatrix} \text{ conjugate } J\text{-lossless and } \hat{D}_{12}(\infty) \text{ nonsingular.}^7$$

4.2. An equivalent problem. Using the special coprime factorization $P = ND^{-1}$ of Lemma 4.1, we show that K is a stabilizing controller for P with $\|\mathcal{F}(P, K)\|_\infty < \gamma$ if and only if K is a stabilizing controller for P_t with $\|\mathcal{F}(P_t, K)\|_\infty < \gamma$ and where P_t is of left-handed stable analytic system type (i.e., an output estimation type problem).

THEOREM 4.3. Suppose that the conditions of Lemma 4.1 hold and $P = ND^{-1}$ is a right coprime factorization as in Lemma 4.1. The following are equivalent:

(1) K is a stabilizing controller for P such that $\|\mathcal{F}(P, K)\|_\infty \leq \gamma$;

(2) K is a stabilizing controller for $P_t \in \mathcal{R}^{(q+m) \times (p+q)}$ such that $\|\mathcal{F}(P_t, K)\|_\infty \leq \gamma$,

where P_t is given by

$$(4.3) \quad P_t = \begin{bmatrix} I_q & 0 \\ N_{21} & N_{22} \end{bmatrix} \begin{bmatrix} 0 & I_p \\ D_{21} & D_{22} \end{bmatrix}^{-1} = \begin{bmatrix} 0 & I_q \\ N_{22} & N_{21} \end{bmatrix} \begin{bmatrix} I_p & 0 \\ D_{22} & D_{21} \end{bmatrix}^{-1}.$$

Proof. Since $\begin{bmatrix} N_{11} & N_{12} \\ D_{11} & D_{12} \end{bmatrix}$ is J -lossless, $D_{12} \in \mathcal{GH}_\infty$ (Lemma 3.3). Now consider the following new factorization of P :

$$(4.4) \quad \begin{bmatrix} \bar{N} \\ \bar{D} \end{bmatrix} = \begin{bmatrix} N \\ D \end{bmatrix} \begin{bmatrix} D_{11} & D_{12} \\ I & 0 \end{bmatrix}^{-1}.$$

⁷ $\hat{N} \in \mathcal{R}^{(m+l) \times (p+q)}$, $\hat{D} \in \mathcal{R}^{(m+l) \times (l+m)}$, so the off-diagonal blocks of \hat{D} are square.

Note that $\bar{D} = \begin{bmatrix} I & 0 \\ \bar{D}_{21} & \bar{D}_{22} \end{bmatrix}$. Hence, by Corollary 2.3, K stabilizes P if and only if $K = \bar{K}_2^{-1} \bar{K}_1$, where

$$(4.5) \quad [\bar{K}_1 \quad \bar{K}_2] \begin{bmatrix} \bar{N}_{21} & -\bar{N}_{22} \\ -\bar{D}_{21} & \bar{D}_{22} \end{bmatrix} = [Q \quad I], \quad \bar{K}_1, \bar{K}_2 \in \mathcal{RH}_\infty, \quad \bar{K}_2(\infty) \text{ nonsingular}$$

and that, in this case,

$$(4.6) \quad \begin{bmatrix} \mathcal{F}(P, K) \\ I \end{bmatrix} = \begin{bmatrix} \bar{N}_{11} & \bar{N}_{12} \\ I & 0 \end{bmatrix} \begin{bmatrix} I \\ Q \end{bmatrix}.$$

Suppose that 1 holds. Let $K = \bar{K}_2^{-1} \bar{K}_1$ be such that \bar{K}_1, \bar{K}_2 satisfy (4.5), with $\mathcal{F}(P, K) \in \gamma\mathcal{BH}_\infty$. Using Lemma 3.4, (4.4), and (4.6), it follows that $I - D_{11}Q \in \mathcal{GH}_\infty$ (equivalently, $I - QD_{11} \in \mathcal{GH}_\infty$) and that $R = (I - QD_{11})^{-1}QD_{12} \in \gamma\mathcal{BH}_\infty$. Now use (4.4) and (4.5) to get

$$(4.7) \quad [K_1 \quad K_2] \begin{bmatrix} N_{22} & -N_{21} \\ -D_{22} & D_{21} \end{bmatrix} = [R \quad I], \quad K_1, K_2 \in \mathcal{RH}_\infty, \quad K_2(\infty) \text{ nonsingular}$$

and

$$(4.8) \quad R \in \gamma\mathcal{BH}_\infty,$$

where

$$[K_1 \quad K_2] = (I - QD_{11})^{-1}[\bar{K}_1 \quad \bar{K}_2].$$

It follows from Corollary 2.3 that $K_2^{-1}K_1 = \bar{K}_2^{-1}\bar{K}_1 = K$ stabilizes P_t , and $\mathcal{F}(P_t, K) = R \in \gamma\mathcal{BH}_\infty$.

Suppose that 2 holds. By Corollary 2.3, let $K = K_2^{-1}K_1$ be such that (4.7) and (4.8) hold ($R = \mathcal{F}(P_t, K)$). Then, by Lemma 3.4, it follows that $D_{11}R + D_{12} \in \mathcal{GH}_\infty$ (equivalently, $I + D_{12}^{-1}D_{11}R \in \mathcal{GH}_\infty$, or $I + RD_{12}^{-1}D_{11} \in \mathcal{GH}_\infty$) and that $(N_{11}R + N_{12}) \times (D_{11}R + D_{12})^{-1} \in \gamma\mathcal{BH}_\infty$. Define

$$\begin{aligned} [\bar{K}_1 \quad \bar{K}_2] &= (I + RD_{12}^{-1}D_{11})^{-1}[K_1 \quad K_2], \\ Q &= (I + RD_{12}^{-1}D_{11})^{-1}RD_{12} = R(D_{11}R + D_{12})^{-1}. \end{aligned}$$

It is easily verified that (4.5) and (4.6) hold, with $\mathcal{F}(P, K) = (N_{11}R + N_{12})(D_{11}R + D_{12})^{-1}$. \square

4.3. The general solution. The necessary condition of Lemma 4.1 reduces the general suboptimal \mathcal{H}_∞ control problem to another suboptimal \mathcal{H}_∞ control problem of left-handed analytic system type (i.e., of output estimation type), by Theorem 4.3. This problem can be solved directly using the theory of § 3.

THEOREM 4.4. $P \in \mathcal{R}^{(l+m) \times (p+q)}$ has an r.c.f. $P = YX^{-1}$ and an l.c.f. $P = \hat{X}^{-1}\hat{Y}$ such that

- (1) $X = \begin{bmatrix} I_p & 0 \\ \hat{X}_{21} & \hat{X}_{22} \end{bmatrix}$, Y_{22}, X_{22} are r.c. and Y_{12} is left invertible in \mathcal{L}_∞ ;
- (2) $\hat{X} = \begin{bmatrix} I_l & \hat{X}_{12} \\ 0 & \hat{X}_{22} \end{bmatrix}$, $\hat{X}_{22}, \hat{Y}_{22}$ are l.c. and \hat{Y}_{21} is right invertible in \mathcal{L}_∞ .

There exists a stabilizing controller K for P such that $\|\mathcal{F}(P, K)\|_\infty < \gamma$ if and only if the following two conditions hold:

$$(4.9) \quad (a) \quad P \text{ has an r.c.f. } P = ND^{-1}, \text{ with } \begin{bmatrix} N_{11} & N_{12} \\ D_{11} & D_{12} \end{bmatrix} \text{ is } J\text{-lossless and } D_{21}(\infty)$$

nonsingular;

(b) P_t has a l.c.f.

$$(4.10) \quad P_t = \hat{D}^{-1} \hat{N}, \quad \text{with} \quad \begin{bmatrix} \hat{N}_{11} & \hat{D}_{11} \\ \hat{N}_{21} & \hat{D}_{21} \end{bmatrix} \text{ conjugate } J\text{-lossless and } \hat{N}_{22}(\infty) \text{ nonsingular},^8$$

where

$$(4.11) \quad P_t = \begin{bmatrix} I_q & 0 \\ N_{21} & N_{22} \end{bmatrix} \begin{bmatrix} 0 & I_p \\ D_{21} & D_{22} \end{bmatrix}^{-1} \in \mathcal{R}^{(q+m) \times (p+q)}.$$

In this case, K is a stabilizing controller for P such that $\|\mathcal{F}(P, K)\|_\infty \leq \gamma$ if and only if $K = \hat{K}_2^{-1} \hat{K}_1$, where

$$(4.12) \quad [\hat{K}_1 \quad \hat{K}_2] = [U \quad I_q] \begin{bmatrix} \hat{D}_{12} & \hat{N}_{12} \\ \hat{D}_{22} & \hat{N}_{22} \end{bmatrix}, \quad U \in \gamma \mathcal{B}\mathcal{H}_\infty^{q \times m}, \quad \hat{K}_2(\infty) \text{ nonsingular}.$$

Equivalently,

$$(4.13) \quad K = \mathcal{F}(K_a, U),$$

$$(4.14) \quad K_a = \begin{bmatrix} \hat{N}_{12} & I_m \\ \hat{N}_{22} & 0 \end{bmatrix}^{-1} \begin{bmatrix} \hat{D}_{12} & 0 \\ \hat{D}_{22} & I_q \end{bmatrix}.$$

Note. $\hat{N}_{22}(\infty)$ nonsingular ensures $\hat{K}_2(\infty)$ is nonsingular for any strictly proper U . It also ensures K_a is proper.

Proof. The result follows from Lemma 4.1, Theorem 4.3, and Theorem 3.9, provided G_t has a right inverse in \mathcal{RL}_∞ . ($G_t = \begin{bmatrix} N_{22} & -N_{21} \\ -D_{22} & D_{21} \end{bmatrix}$). We therefore must show that this is the case, provided P has l.c.f. satisfying 2. This we do as follows.

Since $ND^{-1} = \hat{X}^{-1} \hat{Y}$ are an r.c.f and an l.c.f of P , there exist $U, V, \hat{U}, \hat{V} \in \mathcal{RH}_\infty$ such that

$$(4.15) \quad \begin{bmatrix} D & -U \\ N & V \end{bmatrix} \begin{bmatrix} \hat{V} & \hat{U} \\ -\hat{Y} & \hat{X} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}.$$

Suppose, to obtain a contradiction, that G_t does not have full row rank on ∂ . Then there exist $\lambda \in \partial$ and x_1, x_2 , not both zero, such that

$$\begin{aligned} [x_2 \quad -x_1] G_t(\lambda) &= [0 \quad 0] \\ \Leftrightarrow [0 \quad x_1 \quad 0 \quad x_2] \begin{bmatrix} D & -U \\ N & V \end{bmatrix}(\lambda) &= [0 \quad 0 \quad y_1 \quad y_2] \\ \Leftrightarrow [0 \quad x_1 \quad 0 \quad x_2] &= [0 \quad 0 \quad y_1 \quad y_2] \begin{bmatrix} \hat{V} & \hat{U} \\ -\hat{Y} & \hat{X} \end{bmatrix}(\lambda) \\ \Leftrightarrow y_1 &= 0 \text{ (by the structure of } \hat{X} \text{) and } y_2 [\hat{Y}_{21} \hat{Y}_{22} \hat{X}_{22}](\lambda) = [0 \quad x_1 \quad x_2]. \end{aligned}$$

So G_t has a right inverse in $\mathcal{RL}_\infty \Leftrightarrow \hat{Y}_{21}$ has a right inverse in \mathcal{RL}_∞ . \square

Remark. Suppose that $P(s) = D + C(sI - A)^{-1}B$. Then the factorizations with the requisite triangular structure exist provided (A, B_2, C_2) is stabilizable and detectable. In this case, Y_{12} and \hat{Y}_{21} are, respectively, left invertible and right invertible in \mathcal{RL}_∞ if and only if the matrices

$$\begin{bmatrix} A - \lambda I & B_2 \\ C_1 & D_{12} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} A - \lambda I & B_1 \\ C_2 & D_{21} \end{bmatrix}$$

have full column and row rank, respectively, for all $\lambda \in \partial$. These conditions are also required in other approaches to the \mathcal{H}_∞ synthesis problem [9], [16]–[18], [28].

⁸ $\hat{N} \in \mathcal{R}^{(m+q) \times (p+q)}$, $\hat{D} \in \mathcal{R}^{(m+q) \times (q+m)}$.

Remark. Theorem 4.4 solves the \mathcal{H}_∞ controller synthesis problem in the sense that if and only if we can find the J -lossless coprime factorizations (4.9) and (4.10), the problem has solutions. This is only meaningful if we can actually say something about when we can find these factorizations. In the remainder of the paper, we will develop a state-space answer to the question of when these factorizations exist. However, one advantage of the transfer matrix statement of the result is that *any* method by which J -lossless factorizations can be calculated will be a method by which \mathcal{H}_∞ controllers can be calculated. An alternative to a state-space approach to the calculation of the J -lossless coprime factorizations is a polynomial matrix type approach.

Remark. The necessary condition of Lemma 4.2 plays no apparent role in the necessary and sufficient condition of Theorem 4.4. A number of questions therefore arise. How does the factorization (4.10) imply the existence of the factorization (4.2)? Is it equivalent to (4.2)? If not (and it is not), what additional condition is involved in (4.10)?

A further interesting point is that a dual necessary and sufficient condition, and a dual generator of all \mathcal{H}_∞ controllers, can be given based on Lemma 4.2 as the starting point instead of Lemma 4.1 (this particular route, where the observer is constructed first and the full information feedback second, was taken in [18]). Obviously, since both approaches are necessary and sufficient conditions for the solution of the same problem, they are equivalent. Do they give the same controller generator, however? How precisely are the four J -lossless factorizations involved in the two approaches connected? Some progress has been made toward answering these questions, but a complete answer is yet to emerge. It is true that the generators of all \mathcal{H}_∞ controllers obtained via the two dual approaches are exactly the same. A state space proof of this for the “uncoupled” case of [9] is very straightforward, but the calculations become very involved in the general case. A transfer matrix proof is possible, and will be reported elsewhere.

5. J -spectral and J -lossless factorizations.

5.1. J -spectral factorization.

DEFINITION. A matrix $H \in \mathbb{C}^{2n \times 2n}$ is a Hamiltonian matrix if $\hat{J}H = H^* \hat{J}^*$, $\hat{J} = \begin{bmatrix} 0 & I_n \\ I_n & 0 \end{bmatrix}$ (i.e., $H_{22} = -H_{11}^*$, $H_{12}^* = H_{12}$ and $H_{21}^* = H_{21}$).

If H is Hamiltonian, we say $H \in \text{dom}(\text{Ric})$ if there exists an $n \times n$ matrix Q such that

$$(5.1) \quad H \begin{bmatrix} I \\ Q \end{bmatrix} = \begin{bmatrix} I \\ Q \end{bmatrix} \Lambda, \quad \text{with } \Lambda \mathbb{C}_+ \text{ asymptotically stable.}$$

($\Lambda \mathbb{C}_+$ asymptotically stable means all eigenvalues of Λ have strictly negative real part.) If $H \in \text{dom}(\text{Ric})$, then $Q = \text{Ric}(H)$ has the properties

$$\begin{aligned} Q &= Q^*, \\ (H_{11} + H_{12}Q) &= \Lambda = (-H_{22} + QH_{12})^* \text{ is } \mathbb{C}_+ \text{ asymptotically stable,} \\ QH_{11} + H_{11}^*Q + QH_{12}Q - H_{21} &= 0. \end{aligned}$$

Thus Q is a stabilizing solution to an algebraic Riccati equation.

The equivalence between J -spectral factorization and the solution of indefinite Riccati equations is as follows.

THEOREM 5.1 (see [18]). Suppose that $G \in \mathcal{RH}_\infty^{(l+m) \times (q+m)}(\mathbb{C}_+)$ is given by the realization $G(s) = D + C(sI - A)^{-1}B$, with $A \in \mathbb{C}^{n \times n}$ asymptotically stable.⁹ There exists

⁹ We will also need $G \in \mathcal{RH}_\infty^{(l+m) \times (m+q)}$. The off-diagonal blocks of W are then the square ones instead of the diagonal ones.

a W such that

$$(5.2) \quad \tilde{G}^* J_{lm}(\gamma) G = \tilde{W}^* J_{qm}(\gamma) W \quad \text{with } W \in \mathcal{GH}_\infty^{q+m}(\mathbb{C}_+)$$

if and only if

(1) There exists a nonsingular constant matrix W_∞^{q+m} such that

$$(5.3) \quad D^* J_{lm}(\gamma) D = W_\infty^* J_{qm}(\gamma) W_\infty$$

and

(2) $H \in \text{dom}(\text{Ric})$, where, with $J = J_{lm}(\gamma)$,

$$(5.4) \quad H = \begin{bmatrix} A & 0 \\ -C^* J C & -A^* \end{bmatrix} - \begin{bmatrix} B \\ -C^* J D \end{bmatrix} (D^* J D)^{-1} [D^* J C \quad B^*].$$

In this case, $W \in \mathcal{GH}_\infty(\mathbb{C}_+)$ satisfies (5.2) if and only if, for some solution W_∞ of (5.3), W is given by

$$(5.5a) \quad W(s) = W_\infty + L(sI - A)^{-1} B,$$

where

$$(5.5b) \quad L = J^{-1} W_\infty^{-*} (D^* J C + B^* Q), \quad \text{with } Q = \text{Ric}(H).$$

Proof. For the proof, see [18]. \square

Remark. An alternative statement of the theorem is: If A is asymptotically stable, then there exists a $W \in \mathcal{GH}_\infty(\mathbb{C}_+)$ such that $\tilde{G}^* J G = \tilde{W}^* J W$, $G = D + C(sI - A)^{-1} B$ if and only if there exist $Q = Q^*$, L , W_∞ satisfying the linear matrix equality

$$(5.6) \quad \begin{bmatrix} QA + A^* Q + C^* J C & QB + C^* J D \\ D^* J C + B^* Q & D^* J D \end{bmatrix} = \begin{bmatrix} L^* \\ W_\infty^* \end{bmatrix} J \begin{bmatrix} I & W_\infty \end{bmatrix}$$

and such that

$$(5.7) \quad W_\infty \text{ is nonsingular and } A - BW_\infty^{-1}L \text{ is asymptotically stable.}$$

For discrete time, replace (5.6) with

$$(5.8) \quad \begin{bmatrix} A & B \\ C & D \end{bmatrix}^* \begin{bmatrix} Q & 0 \\ 0 & J \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} I & O \\ L & W_\infty \end{bmatrix}^* \begin{bmatrix} Q & 0 \\ 0 & J \end{bmatrix} \begin{bmatrix} I & 0 \\ L & W_\infty \end{bmatrix}.$$

This can be proved either directly or from the continuous time result via bilinear transformation of the realizations—see, e.g., [14, § 2.2]. Note that, if A is \mathbb{D}_+ asymptotically stable, then $(A - I)$ is nonsingular, which is required to transform the realization.

To calculate Q , L , W_∞ we must obtain the stable eigenspace of the Hamiltonian matrix (for continuous time). For discrete time, an analogous algorithm involving a symplectic matrix is appropriate [31].

Note also that more conventional-looking forms of the Riccati equations may be obtained from (5.6) and (5.8) by substituting for L and W_∞ from the (2,1) and (2,2) blocks into the (1,1) block. (Equivalently, take Schur complements with respect to the (2,2) block.) \square

The following lemma is useful in state-space calculations, especially in § 6.

LEMMA 5.2. Suppose that D_2 is a nonsingular matrix. Let

$$G_1 \stackrel{s}{=} \begin{bmatrix} A & B \\ C_1 & D_1 \end{bmatrix}$$

and

$$G_2 \stackrel{s}{=} \begin{bmatrix} A - BD_2^{-1}C_2 & BD_2^{-1} \\ C_1 - D_1D_2^{-1}C_2 & D_1D_2^{-1} \end{bmatrix} = \begin{bmatrix} A & B \\ C_1 & D_1 \end{bmatrix} \begin{bmatrix} I & 0 \\ C_2 & D_2 \end{bmatrix}^{-1}.$$

The Hamiltonian matrices H_1 and H_2 associated via (5.4) with the realizations above are identical. Furthermore, in obvious notation,

$$\begin{bmatrix} L_2 & W_2 \end{bmatrix} = \begin{bmatrix} L_1 & W_1 \end{bmatrix} \begin{bmatrix} I & 0 \\ C_2 & D_2 \end{bmatrix}^{-1}.$$

Proof. The proof is obtained by direct calculation. This is also easily seen from (5.6) (or (5.8) for the discrete time analogue). \square

5.2. J -lossless factorization. We now consider when the factorization $\tilde{G}^* J G = \tilde{W}^* J W$ is such that $G W^{-1}$ is J -lossless. We will show that this is the case if and only if the stabilizing solution to the algebraic Riccati equation is positive semidefinite.

THEOREM 5.3. *Let $G(s) = D + C(sI - A)^{-1}B \in \mathcal{RH}_\infty^{(l+m) \times (q+m)}(\mathbb{C}_+)$, with $A \in \mathbb{C}_+$ asymptotically stable. Then there exists a $W \in \mathcal{GH}_\infty^{q+m}(\mathbb{C}_+)$ such that $G W^{-1}$ is J -lossless if and only if*

(1) *There exists a nonsingular $W_\infty \in \mathbb{C}^{(q+m) \times (q+m)}$ such that $W_\infty^* J_{qm}(\gamma) W_\infty = D^* J_{lm}(\gamma) D$;*

(2) *$H \in \text{dom}(\text{Ric})$, with H as in (5.4);*

(3) *$\text{Ric}(H) \geq 0$.*

Proof. Conditions 1 and 2 are necessary and sufficient for the existence of a $W \in \mathcal{GH}_\infty$ such that $(G W^{-1})^* J (G W^{-1}) = J$ on $j\mathbb{R}$. Using (5.5), obtain a realization of $X = G W^{-1}$;

$$(5.9) \quad X = G W^{-1} \stackrel{s}{=} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} I & 0 \\ L & W_\infty \end{bmatrix}^{-1} = \begin{bmatrix} \bar{A} & \bar{B} \\ \bar{C} & \bar{D} \end{bmatrix}.$$

Let $Q = \text{Ric}(H)$. Using Lemma 5.2, it follows from (5.9) and (5.6) that

$$(5.10) \quad \bar{D}^* J \bar{D} = (D W_\infty^{-1})^* J (D W_\infty^{-1}) = J,$$

$$(5.11) \quad \bar{D}^* J \bar{C} + \bar{B}^* Q = 0,$$

$$(5.12) \quad Q \bar{A} + \bar{A}^* Q + \bar{C}^* J \bar{C} = 0.$$

Since \bar{A} is asymptotically stable, $X \in \mathcal{RH}_\infty$. By Lemma 3.3, X is J -lossless $\Leftrightarrow X_{22} \in \mathcal{GH}_\infty$. Partition \bar{C} , \bar{B} , \bar{D} appropriately. By the asymptotic stability of \bar{A} , $X_{22} \in \mathcal{GH}_\infty \Leftrightarrow \bar{A} - \bar{B}_2 \bar{D}_{22}^{-1} \bar{C}_2$ is asymptotically stable. We now show this is the case if and only if $Q \geq 0$.

From the (2,2) block of (5.10), we have $\bar{D}_{12}^* \bar{D}_{12} - \gamma^2 \bar{D}_{22}^* \bar{D}_{22} = -\gamma^2 I$, so

$$(5.13) \quad \bar{D}_{22} \text{ is nonsingular, and } \|\bar{D}_{12} \bar{D}_{22}^{-1}\| < \gamma.$$

Also, from the (2,1) block of (5.11), we have that

$$(5.14) \quad \bar{D}_{12}^* \bar{C}_1 - \gamma^2 \bar{D}_{22}^* \bar{C}_2 + \bar{B}_2^* Q = 0.$$

Hence

$$\begin{aligned} & Q(\bar{A} - \bar{B}_2 \bar{D}_{22}^{-1} \bar{C}_2) + (\bar{A} - \bar{B}_2 \bar{D}_{22}^{-1} \bar{C}_2)^* Q \\ &= -\bar{C}_1^* \bar{C}_1 + \gamma^2 \bar{C}_2^* \bar{C}_2 + \bar{C}_2^* \bar{D}_{22}^{-*} [\bar{D}_{12}^* \bar{C}_1 - \gamma^2 \bar{D}_{22}^* \bar{C}_2] \\ & \quad + [\bar{D}_{12}^* \bar{C}_1 - \gamma^2 \bar{D}_{22}^* \bar{C}_2]^* \bar{D}_{22}^{-1} \bar{C}_2, \text{ using (5.14) and (5.12)} \\ &= -[\bar{C}_1^* \bar{C}_2^*] \begin{bmatrix} I & -\bar{D}_{12} \bar{D}_{22}^{-1} \\ -\bar{D}_{22}^* \bar{D}_{12}^* & \gamma^2 I \end{bmatrix} \begin{bmatrix} \bar{C}_1 \\ \bar{C}_2 \end{bmatrix} = -\bar{C}^* R \bar{C}. \end{aligned}$$

Since $\|\bar{D}_{12}\bar{D}_{22}^{-1}\| < \gamma$, $R > 0$ and since \bar{A} is asymptotically stable, $(R^{1/2}\bar{C}, \bar{A} - \bar{B}_2\bar{D}_{22}^{-1}\bar{C}_2)$ is detectable. Therefore $\bar{A} - \bar{B}_2\bar{D}_{22}^{-1}\bar{C}_2$ is asymptotically stable if and only if $Q \geq 0$ [34]. \square

Remark. For discrete time, the result is the following.

Let $G(z) = D + C(zI - A)^{-1}B \in \mathcal{RH}_\infty^{(l+m) \times (q+m)}(\mathbb{D}_+)$, with A asymptotically stable. There exists a $W \in \mathcal{GH}_\infty^{q+m}(\mathbb{D}_+)$ such that GW^{-1} is J -lossless if and only if there exists $Q \geq 0$, L , W_∞ satisfying (5.7) and (5.8).

The sufficiency proof is easy: By direct calculation, $(GW^{-1})^*J(GW^{-1}) = J + (1 - |z|^2)\bar{B}(\bar{z}I - \bar{A}^*)^{-1}Q(zI - \bar{A})^{-1}\bar{B}$.

For necessity, note that $GW^{-1} \in \mathcal{RH}_\infty$ J -lossless $\Rightarrow (GW^{-1})_{22} \in \mathcal{GH}_\infty \Rightarrow \bar{D}_{22} = (DW_\infty^{-1})_{22}(\infty)$ is nonsingular. The proof then goes similarly to the continuous case and can be compactly formulated as follows.

First, note that

$$(5.15) \quad \begin{bmatrix} \bar{A} & \bar{B} \\ \bar{C} & \bar{D} \end{bmatrix}^* \begin{bmatrix} Q & 0 \\ 0 & J \end{bmatrix} \begin{bmatrix} \bar{A} & \bar{B} \\ \bar{C} & \bar{D} \end{bmatrix} = \begin{bmatrix} Q & 0 \\ 0 & J \end{bmatrix}.$$

Partition \bar{B} , \bar{C} , \bar{D} and consider $(X^{-1})^*(5.15)X^{-1}$, with

$$X = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ \bar{C}_2 & \bar{D}_{21} & \bar{D}_{22} \end{bmatrix}.$$

Examining the (1, 1) block, we obtain a Lyapunov equation for $\bar{A} - \bar{B}_2\bar{D}_{22}^{-1}\bar{C}_2$. \square

6. State-space formulae. We now calculate the controller generator of Theorem 4.4 by applying Theorem 5.3 twice. To avoid clutter, we will not specify the dimensions of matrices¹⁰. We will also use J instead of $J_{pq}(\gamma)$, etc. The disadvantage of this is that the symbol J is used within the same equation to mean different matrices. No confusion should however occur: a J that multiplies a B , C , or D matrix is different (of larger dimension) from one that multiplies a W , L , or M matrix.

The results will be stated for the continuous case, with remarks giving the changes for discrete time—the actual formula for the generator of all controllers is exactly the same, with the obvious replacement of continuous time L 's and W_∞ 's, etc., with their discrete time counterparts (see also [28]).

6.1. J -lossless coprime factorizations. The following lemma facilitates calculations involving matrix fractions.

LEMMA 6.1. Suppose that D_2 and D_3 are nonsingular matrices, and

$$\begin{bmatrix} N \\ D \end{bmatrix} \stackrel{s}{=} \begin{bmatrix} A & B \\ C_1 & D_1 \\ C_2 & D_2 \end{bmatrix} \begin{bmatrix} I & 0 \\ F & D_3 \end{bmatrix}^{-1}.$$

Then

$$ND^{-1} \stackrel{s}{=} \begin{bmatrix} A & B \\ C_1 & D_1 \end{bmatrix} \begin{bmatrix} I & 0 \\ C_2 & D_2 \end{bmatrix}^{-1}.$$

Proof. The proof is obtained by direct calculation. \square

THEOREM 6.2. Let P have state-space realization $D + C(sI - A)^{-1}B$, with (A, B) stabilizable and (C, A) detectable. Define \bar{C} and \bar{D} to be the matrices

$$(6.1) \quad \bar{D} = \begin{bmatrix} D_{11} & D_{12} \\ I & 0 \end{bmatrix}, \quad \bar{C} = \begin{bmatrix} C_1 \\ 0 \end{bmatrix}.$$

¹⁰ See the theorem statements in § 4 to get explicit dimensions.

Then P has an r.c.f. such that

$$(6.2) \quad P = ND^{-1}, \quad N, D, \text{ r.c., and } \begin{bmatrix} N_{11} & N_{12} \\ D_{11} & D_{12} \end{bmatrix} \text{ } J\text{-lossless and } D_{21}(\infty) \text{ nonsingular}$$

if and only if

(1) There exists a nonsingular matrix W_∞ such that

$$(6.3) \quad W_\infty^* J W_\infty = \bar{D}^* J \bar{D},$$

equivalently,

$$(6.4) \quad D_{12}^* D_{12} > 0 \quad \text{and} \quad \Phi = D_{11}^* [I - D_{12} (D_{12}^* D_{12})^{-1} D_{12}^*] D_{11} < \gamma^2 I;$$

(2) $H_X \in \text{dom}(\text{Ric})$, where

$$(6.5) \quad H_X = \begin{bmatrix} A & 0 \\ -\bar{C}^* J \bar{C} & -A^* \end{bmatrix} - \begin{bmatrix} B \\ -\bar{C}^* J \bar{D} \end{bmatrix} (\bar{D}^* J \bar{D})^{-1} [\bar{D}^* J \bar{C} \quad B^*];$$

(3) $X_\infty = \text{Ric}(H_X) \geq 0$.

In this case, N and D are given by the realizations

$$(6.6) \quad \begin{bmatrix} N \\ D \end{bmatrix} \stackrel{s}{=} \begin{bmatrix} A & B \\ C & D \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ L & W_\infty \end{bmatrix}^{-1},$$

where W_∞ is any solution to (6.3) and

$$(6.7) \quad W_\infty^* J L = \bar{D}^* J \bar{C} + B^* X_\infty.$$

One solution to (6.3) is given by

$$(6.8) \quad W_\infty = \begin{bmatrix} (D_{12}^* D_{12})^{-1/2} D_{12}^* D_{11} & (D_{12}^* D_{12})^{1/2} \\ \gamma^{-1} (\gamma^2 I - \Phi)^{1/2} & 0 \end{bmatrix}.$$

Proof. Let F be any matrix such that $A - BF$ is asymptotically stable and define Y and X by

$$(6.9) \quad \begin{bmatrix} Y \\ X \end{bmatrix} \stackrel{s}{=} \begin{bmatrix} A & B \\ C & D \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ F & I \end{bmatrix}^{-1}.$$

Note that Y, X are r.c. since (C, A) is detectable. Define G by

$$(6.10) \quad G = \begin{bmatrix} Y_{11} & Y_{12} \\ X_{11} & X_{12} \end{bmatrix} \stackrel{s}{=} \begin{bmatrix} A & B \\ \bar{C} & \bar{D} \end{bmatrix} \begin{bmatrix} I & 0 \\ F & I \end{bmatrix}^{-1}.$$

The coprime factorization in (6.2) exists if and only if there exists a $W \in \mathcal{GH}_\infty$ such that $N = YW^{-1}$, $D = XW^{-1}$, and GW^{-1} is J -lossless. The result now follows from Theorem 5.3, using Lemma 5.2.

That (6.8) is a solution to (6.3) follows by direct calculation. \square

Remark. For discrete time, X_∞, L, W_∞ must satisfy

$$\begin{bmatrix} A & B \\ \bar{C} & \bar{D} \end{bmatrix}^* \begin{bmatrix} X_\infty & 0 \\ 0 & J \end{bmatrix} \begin{bmatrix} A & B \\ \bar{C} & \bar{D} \end{bmatrix} = \begin{bmatrix} I & 0 \\ L & W_\infty \end{bmatrix}^* \begin{bmatrix} X_\infty & 0 \\ 0 & J \end{bmatrix} \begin{bmatrix} I & 0 \\ L & W_\infty \end{bmatrix},$$

W_∞ nonsingular, $A - BW_\infty^{-1}L$ asymptotically stable, and $X_\infty \geq 0$.

Note that conditions analogous to (6.4) and a formula for W_∞ of the form (6.8) can be simply obtained for the discrete case by considering

$$\tilde{D}_{11} = \begin{bmatrix} X_\infty^{1/2} B_1 \\ D_{11} \end{bmatrix}, \quad \tilde{D}_{12} = \begin{bmatrix} X_\infty^{1/2} B_2 \\ D_{12} \end{bmatrix}.$$

Remark. The Riccati equations obtained from Theorem 6.2 are associated with the full information or state feedback \mathcal{H}_∞ synthesis problem [9], [27]. They also arise in the theory of linear quadratic differential games, which can be used as the basis for a solution to the \mathcal{H}_∞ synthesis problem [27], [28].

Remark. In § 4.1, we saw that

$$\begin{bmatrix} 0 \\ w - w_{\text{worst}} \end{bmatrix} = D^{-1} \begin{bmatrix} w \\ u_{\text{opt}} \end{bmatrix}.$$

From (6.6), $D^{-1} \stackrel{s}{=} \begin{bmatrix} A & B \\ L & w_\infty \end{bmatrix}$. Hence

$$0 = L_1 x + W_{11} w + W_{12} u_{\text{opt}},$$

giving

$$u_{\text{opt}} = -W_{12}^{-1}(L_1 x + W_{11} w).$$

Note that u_{opt} is *not* a pure state feedback in general, which is the reason for the term “full information” introduced in [9]. The condition for pure state feedback is $W_{11} = 0$, which can be mapped back to original data from (6.8): $D_{12}^* D_{11} = 0$ in continuous time, and $B_2^* X_\infty B_1 + D_{12}^* D_{11} = 0$ in discrete time. That is, the contributions of the disturbance w and the control u to the controlled variable z are independent. In this case, the second condition in (6.4) becomes $D_{11}^* D_{11} < \gamma^2 I$ in continuous time and $D_{11}^* D_{11} + B_1^* X_\infty B_1 < \gamma^2 I$ in discrete time.

Similarly, from

$$\begin{bmatrix} u - u_{\text{opt}} \\ 0 \end{bmatrix} = D^{-1} \begin{bmatrix} w_{\text{worst}} \\ u \end{bmatrix},$$

we have

$$0 = L_2 x + W_{21} w_{\text{worst}},$$

so

$$w_{\text{worst}} = -W_{21}^{-1} L_2 x.$$

Note that these formulae are valid for both continuous time and discrete time systems—we just need to use the appropriate L ’s and W ’s.

THEOREM 6.3. *Let P have state-space realization $D + C(sI - A)^{-1}B$, with (A, B) stabilizable and (C, A) detectable. Define \bar{B} and \bar{D} to be the matrices*

$$(6.11) \quad \bar{D} = \begin{bmatrix} D_{11} & I \\ D_{21} & 0 \end{bmatrix}, \quad \bar{B} = \begin{bmatrix} B_1 & 0 \end{bmatrix}.$$

Then P has an l.c.f. such that

$$(6.12) \quad P = \hat{D}^{-1} \hat{N}, \quad \hat{D}, \hat{N} \text{ l.c. and } \begin{bmatrix} \hat{N}_{11} & \hat{D}_{11} \\ \hat{N}_{21} & \hat{D}_{21} \end{bmatrix} \text{ conjugate } J\text{-lossless and } \hat{D}_{12}(\infty) \text{ nonsingular}$$

if and only if

(1) *There exists a nonsingular matrix W_∞ such that*

$$(6.13) \quad \hat{W}_\infty J \hat{W}_\infty^* = \bar{D} J \bar{D}^*,$$

equivalently,

$$(6.14) \quad D_{21} D_{21}^* > 0 \quad \text{and} \quad \hat{\Phi} = D_{11} [I - D_{21}^* (D_{21} D_{21}^*)^{-1} D_{21}] D_{11}^* < \gamma^2 I;$$

(2) $H_Y \in \text{dom}(\text{Ric})$, where

$$(6.15) \quad H_Y = \begin{bmatrix} A^* & 0 \\ -\bar{B} J \bar{B}^* & -A \end{bmatrix} - \begin{bmatrix} C^* \\ -\bar{B} J \bar{D}^* \end{bmatrix} (\bar{D} J \bar{D}^*)^{-1} [\bar{D} J \bar{B}^* \quad C];$$

(3) $Y_\infty = \text{Ric}(H) \geq 0$.

In this case, \hat{N} and \hat{D} are given by the realizations

$$(6.16) \quad [\hat{N} \quad \hat{D}] \stackrel{s}{=} \begin{bmatrix} I & M \\ 0 & \hat{W}_\infty \end{bmatrix}^{-1} \begin{bmatrix} A & B & 0 \\ C & D & I \end{bmatrix},$$

where \hat{W}_∞ is any solution to (6.13) and

$$(6.17) \quad MJ\hat{W}_\infty^* = \bar{B}J\bar{D}^* + Y_\infty C^*.$$

One solution to (6.13) is given by

$$(6.18) \quad \hat{W}_\infty = \begin{bmatrix} D_{11}D_{21}^*(D_{21}D_{21}^*)^{-1/2} & \gamma^{-1}(\gamma^2 I - \hat{\Phi})^{1/2} \\ (D_{21}D_{21}^*)^{1/2} & 0 \end{bmatrix}.$$

Remark. For discrete time, Y_∞ , M , \hat{W}_∞ must satisfy

$$\begin{bmatrix} A & \bar{B} \\ C & \bar{D} \end{bmatrix} \begin{bmatrix} Y_\infty & 0 \\ 0 & J \end{bmatrix} \begin{bmatrix} A & \bar{B} \\ C & \bar{D} \end{bmatrix}^* = \begin{bmatrix} I & M \\ 0 & \hat{W}_\infty \end{bmatrix} \begin{bmatrix} Y_\infty & 0 \\ 0 & J \end{bmatrix} \begin{bmatrix} I & M \\ 0 & \hat{W}_\infty \end{bmatrix}^*,$$

\hat{W}_∞ nonsingular, $A - M\hat{W}_\infty^{-1}C$ asymptotically stable, and $Y_\infty \geq 0$.

6.2. Calculation of P_t .

LEMMA 6.4 (Hypotheses as for Theorem 6.2). Suppose also that 1, 2, 3 of Theorem 6.2 hold, and that W_∞ is given by (6.8). For convenience, write

$$(6.19) \quad W_\infty = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & 0 \end{bmatrix}.$$

Then

$$(6.20) \quad P_t = \begin{bmatrix} I & 0 \\ N_{21} & N_{22} \end{bmatrix} \begin{bmatrix} 0 & I \\ D_{21} & D_{22} \end{bmatrix}^{-1}$$

is given by the realization

$$(6.21) \quad P_t \stackrel{s}{=} \begin{bmatrix} A_t & B_t \\ C_t & D_t \end{bmatrix} = \begin{bmatrix} A & B_1 & B_2 \\ L_1 & W_{11} & W_{12} \\ C_2 & D_{21} & D_{22} \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ L_2 & W_{21} & 0 \\ 0 & 0 & I \end{bmatrix}^{-1}.$$

Proof. The proof follows from the realization (6.6) for N , D , and Lemma 6.1, noting that

$$\begin{array}{ccc|ccc} I & 0 & & A & B_1 & B_2 \\ N_{21} & N_{22} & & L_1 & W_{11} & W_{12} \\ & & & C_2 & D_{21} & D_{22} \\ 0 & I & \stackrel{s}{=} & L_2 & W_{21} & 0 \\ D_{21} & D_{22} & & 0 & 0 & I \end{array} \begin{bmatrix} I & 0 \\ L & W_\infty \end{bmatrix}^{-1}.$$

□

6.3. Factorization of P_t and the controller generator.

THEOREM 6.5 (Hypotheses and notation as for Lemma 6.4). Define \bar{B}_t and \bar{D}_t to be the matrices

$$(6.33) \quad \begin{bmatrix} \bar{B}_t \\ \bar{D}_t \end{bmatrix} = \begin{bmatrix} B_1 & 0 \\ W_{11} & I \\ D_{21} & 0 \end{bmatrix} \begin{bmatrix} W_{21} & 0 \\ 0 & I \end{bmatrix}^{-1}.$$

Then P_t has an l.c.f. such that

$$(6.24) \quad P = \hat{D}^{-1} \hat{N}, \hat{D}, \hat{N} \text{ l.c. and } \begin{bmatrix} \hat{N}_{11} & \hat{D}_{11} \\ \hat{N}_{21} & \hat{D}_{21} \end{bmatrix} \text{ conjugate } J\text{-lossless}$$

if and only if

(1) There exists a nonsingular matrix \tilde{W}_∞ such that

$$(6.25) \quad \tilde{W}_\infty J \tilde{W}_\infty^* = \bar{D}_t J \bar{D}_t^*;$$

(2) $H_Z \in \text{dom}(\text{Ric})$, where

$$(6.26) \quad H_Z = \begin{bmatrix} A_t^* & 0 \\ -\bar{B}_t J \bar{B}_t^* & -A_t \end{bmatrix} - \begin{bmatrix} C_t^* \\ -\bar{B}_t J \bar{D}_t^* \end{bmatrix} (\bar{D}_t J \bar{D}_t^*)^{-1} [\bar{D}_t J \bar{B}_t^* \quad C_t];$$

(3) $Z_\infty = \text{Ric}(H) \geq 0$.

In this case, \hat{N} and \hat{D} are given by the realizations

$$(6.27) \quad [\hat{N} \quad \hat{D}] \stackrel{s}{=} \begin{bmatrix} I & \tilde{M} \\ 0 & \tilde{W}_\infty \end{bmatrix}^{-1} \begin{bmatrix} A_t & B_t & 0 \\ C_t & D_t & I \end{bmatrix},$$

where \tilde{W}_∞ is any solution to (6.25) and

$$(6.28) \quad \tilde{M} J \tilde{W}_\infty^* = \bar{B}_t J \bar{D}_t^* + Z_\infty C_t^*.$$

Also, the controller generator K_a (see (4.14)) is given by

$$(6.29) \quad K_a = \begin{bmatrix} \hat{N}_{12} & I \\ \hat{N}_{22} & 0 \end{bmatrix}^{-1} \begin{bmatrix} \hat{D}_{12} & 0 \\ \hat{D}_{22} & I \end{bmatrix}$$

$$(6.30) \quad \stackrel{s}{=} \begin{bmatrix} I & B_2 & \tilde{M}_1 \\ 0 & W_{12} & \tilde{W}_{11} \\ 0 & D_{22} & \tilde{W}_{21} \end{bmatrix}^{-1} \begin{bmatrix} A_t & 0 & \tilde{M}_2 \\ C_{t1} & 0 & \tilde{W}_{12} \\ C_{t2} & I & \tilde{W}_{22} \end{bmatrix}.$$

Proof. The proof follows from the realization (6.21) for P_t and Theorem 6.3. Formula (6.30) for K_a is obtained from (6.27) by using the dual of Lemma 6.1. \square

Remark. Theorem 4.4 requires us to ensure that $\tilde{N}_{22}(\infty)$ is nonsingular, which is equivalent to $\begin{bmatrix} W_{12} & \tilde{W}_{11} \\ D_{22} & \tilde{W}_{21} \end{bmatrix}$ nonsingular. This is always true if $D_{22} = 0$.

Remark. For discrete time, Z_∞ , \tilde{M} , \tilde{W}_∞ must satisfy

$$\begin{bmatrix} A_t & \bar{B}_t \\ C_t & \bar{D}_t \end{bmatrix} \begin{bmatrix} Z_\infty & 0 \\ 0 & J \end{bmatrix} \begin{bmatrix} A_t & \bar{B}_t \\ C_t & \bar{D}_t \end{bmatrix}^* = \begin{bmatrix} I & \tilde{M} \\ 0 & \tilde{W}_\infty \end{bmatrix} \begin{bmatrix} Z_\infty & 0 \\ 0 & J \end{bmatrix} \begin{bmatrix} I & \tilde{M} \\ 0 & \tilde{W}_\infty \end{bmatrix}^*,$$

\tilde{W}_∞ nonsingular, $A_t - \tilde{M} \tilde{W}_\infty^{-1} C_t$ asymptotically stable, and $Z_\infty \geq 0$.

In [28] these formulae are derived using a state-space approach to the discrete time problem based on game theoretic methods.

Remark. It is an interesting exercise to carry out the calculations in this section for the special case considered in [9]—set $D_{11} = 0$, $D_{22} = 0$, $D_{12}^* [C_1 \quad D_{12}] = [0 \quad I]$, and $D_{21} [B_1^* \quad D_{21}^*] = [0 \quad I]$.

Remark. For (6.30) to be a generator of all \mathcal{H}_∞ controllers for P , the assumptions of Theorem 4.4 must be satisfied. The state-space versions of these assumptions, namely that (A, B_2, C_2) is stabilizable and detectable and $\begin{bmatrix} A - \lambda I & B_2 \\ C_1 & D_{12} \end{bmatrix}$ and $\begin{bmatrix} A - \lambda I & B_1 \\ C_2 & D_{21} \end{bmatrix}$ have full column and row rank, respectively, on ∂ , are necessary conditions for X_∞ , Y_∞ , Z_∞ of Theorems 6.2, 6.3, and 6.5 to exist. The (A, B_2, C_2) stabilizable/detectable condition is a necessary condition for the existence of an internally stabilizing controller, but the full rank on the boundary conditions are not necessary for the existence of an \mathcal{H}_∞ controller. If these conditions do not hold, the results say nothing about whether a controller exists and certainly cannot calculate one.

Remark. It is known (see [9], [17], [18], [27]) that the Riccati equation solutions involved in the factorizations of Theorems 6.2, 6.3, and 6.5 are related by $Z_\infty = Y_\infty(I - \gamma^{-1}X_\infty Y_\infty)^{-1}$. Although this suggests an elegant transfer matrix level connection between these factorizations, none has emerged so far. \square

7. Conclusion. In this paper the fundamental relationship between \mathcal{H}_∞ controller synthesis via the technique pioneered in [9] and analytic systems has been elucidated. This both clarifies the structure recognized in [9], exposing the role played by internal stability, and enables transfer function, or operator theoretic, methods to be employed. The result is that the solutions to the (suboptimal) \mathcal{H}_∞ control problem exists if and only if two coupled J -lossless coprime factorizations exist. The second coprime factorization provides a linear fractional representation for all solutions.

Thus, except for the final state-space calculations, the results are independent of whether continuous time or discrete time systems are considered. The state-space formulae can be developed in a few pages, and their structure is exactly the same, whether in continuous or discrete time.

The approach would not, however, appear to offer an easy extension to the optimal case, where more general factorizations need to be considered. Direct state-space methods using descriptor systems seem to offer the easiest approach to the optimal case [17], [30].

Acknowledgment. It is my pleasure to acknowledge the contribution made by David Limebeer to the development of the work in this paper. In particular, we have worked on several approaches to the discrete-time problem, and I am grateful for his permission to include material here resulting from collaborative efforts.

REFERENCES

- [1] V. M. ADAMJAN, D. Z. AROV, AND M. G. KREIN, *Infinite block Hankel matrices and related extension problems*, Amer. Math. Soc. Trans. Ser. 2, 111 (1978), pp. 133–156.
- [2] J. A. BALL AND N. COHEN, *Sensitivity minimization in an H^∞ norm: Parametrization of all sub-optimal solutions*, Internat. J. Control, 46 (1987), pp. 785–816.
- [3] J. A. BALL AND J. W. HELTON, *A Beurling–Lax theorem for the Lie group $U(m, n)$ which contains most classical interpolation theory*, J. Oper. Theory, 9 (1983), pp. 107–142.
- [4] ———, *Shift invariant subspaces, passivity, reproducing kernels and \mathcal{H}^∞ -optimization*, in Operator Theory: Advances and Applications, 35, Birkhäuser-Verlag, Basel, 1988.
- [5] J. A. BALL AND A. C. M. RAN, *Optimal Hankel norm model reductions and Weiner–Hopf factorization, I: The canonical case*, SIAM J. Control Optim., 25 (1987), pp. 362–383.
- [6] H. BART, I. Gohberg, AND M. A. KAASHOEK, *Minimal Factorization of Matrix and Operator Functions*, Birkhäuser-Verlag, Basel, 1979.
- [7] D. S. BERNSTEIN AND W. M. HADDAD, *LQG control with an H_∞ performance bound: A Riccati equation approach*, IEEE Trans. Automat. Control., submitted.
- [8] J. C. DOYLE et al., *Lecture notes in advances in multivariable control*, ONR/Honeywell Workshop, Minneapolis, MN, 1984.
- [9] J. C. DOYLE, K. GLOVER, P. KHARGONEKAR, AND B. FRANCIS, *State-space solutions to standard H_2 and H_∞ control problems*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 831–847.
- [10] H. DYM, *J. contractive matrix functions, reproducing kernel Hilbert spaces and interpolation*, CBMS Regional Conference Series in Mathematics, Vol. 71, Amer. Math. Soc., Providence, RI, 1989.
- [11] B. A. FRANCIS, *A Course in H_∞ Control Theory*, Lecture Notes in Control and Information Sciences 88, Springer-Verlag, Berlin, New York, 1987.
- [12] B. A. FRANCIS AND J. C. DOYLE, *Linear control theory with an \mathcal{H}_∞ optimality criterion*, SIAM J. Control Optim., 25 (1987), pp. 815–844.
- [13] Y. GENIN, P. VAN DOOREN, AND T. KAILATH, *On Σ -lossless transfer functions and related questions*, Linear Algebra Appl., 50, (1984), pp. 251–1193.
- [14] K. GLOVER, *All optimal Hankel-norm approximations of linear multivariable systems and their L^∞ -error bounds*, Internat. J. Control, 39 (1984), pp. 1115–1193.

- [15] K. GLOVER, *Minimum entropy and risk sensitive control: The continuous time case*, in Proc. 28th IEEE C.D.C., Tampa, FL, 1989, pp. 383–391.
- [16] K. GLOVER AND J. C. DOYLE, *State-space formulae for all stabilizing controllers that satisfy a H^∞ norm bound and relations to risk sensitivity*, Systems Control Lett., 11, (1988), pp. 167–172.
- [17] K. GLOVER, D. J. N. LIMEBEER, J. DOYLE, E. M. KASENALLY, AND M. G. SAFONOV, *A characterization of all the solutions to the four block general distance problem*, SIAM J. Control Optim., 29 (1991), pp. 283–324.
- [18] M. GREEN, K. GLOVER, D. J. N. LIMEBEER, AND J. DOYLE, *A J -spectral factorization approach to \mathcal{H}_∞ control*, SIAM J. Control Optim., 28 (1990), pp. 1350–1371.
- [19] J. W. HELTON ET AL., *Operator theory, analytic functions, matrices and electrical engineering*, Conference Board of the Mathematical Sciences, Regional Conference Series in Mathematics 68, American Mathematical Society, Providence, RI, 1987.
- [20] Y. S. HUNG, *\mathcal{RH}_∞ control—Part I: model matching,—part II: Solution for controllers*, Internat. J. Control, 49 (1989), pp. 1291–1330; pp. 1331–1359.
- [21] P. P. KHARGONEKAR, I. R. PETERSEN, AND M. A. ROTE, *H_∞ optimal control with state feedback*, IEEE Trans. Automat. Control, 33 (1988), pp. 783–786.
- [22] P. P. KHARGONEKAR, I. R. PETERSEN, AND K. ZHOU, *Robust stabilization of uncertain linear systems: Quadratic stabilizability and \mathcal{H}_∞ control theory*, IEEE Trans. Automat. Control, 35 (1990), pp. 356–361.
- [23] H. KIMURA, *Directional interpolation approach to \mathcal{H}^∞ -optimization and robust control*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 1085–1093.
- [24] H. KIMURA AND R. KAWATANI, *Synthesis of H^∞ controllers based on conjugation*, in Proc. IEEE C.D.C., 1988.
- [25] H. KWAKERNAAK, *A polynomial approach to minimax frequency domain optimization of multivariable feedback systems*, Internat. J. Control, 44 (1986), pp. 117–156.
- [26] D. J. N. LIMEBEER AND B. D. O. ANDERSON, *An interpolation theory approach to H^∞ controller degree bounds*, Linear Algebra Appl., 98 (1988), pp. 347–386.
- [27] D. J. N. LIMEBEER, B. D. O. ANDERSON, P. P. KHARGONEKAR, AND M. GREEN, *A game theoretic approach to \mathcal{H}_∞ control for time-varying systems*, SIAM J. Control Optim., 30 (1992), pp. 262–283.
- [28] D. J. N. LIMEBEER, M. GREEN, AND D. WALKER, *Discrete time H^∞ control*, 28th IEEE C.D.C., Tampa, FL, December 1989.
- [29] D. J. N. LIMEBEER AND G. D. HALIKIAS, *An analysis of pole zero cancellations in H^∞ optimal control problems of the second kind*, SIAM J. Control Optim., 26 (1988), pp. 646–677.
- [30] D. J. N. LIMEBEER, E. M. KASENALLY, M. G. SAFONOV, AND I. JAIMOUKA, *A characterization of all the solutions to the four block general distance problem*, in Proc. IEEE C.D.C., 1988.
- [31] T. PAPPAS, A. J. LAUB, AND N. R. SAUDELL, JR., *On the numerical solution of the discrete-time algebraic Riccati equation*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 631–641.
- [32] I. R. PETERSEN, *Disturbance attenuation and \mathcal{H}_∞ optimization: A design method based on the algebraic Riccati equation*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 427–429.
- [33] G. TADMOR, *Worst-case design in the time domain. The maximum principle and the standard \mathcal{H}_∞ -problem*, Math. Control Signals Systems, 3 (1990), pp. 301–324.
- [34] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, New York, 1979.
- [35] M. VIDYASAGAR, *Control Systems Synthesis: A Factorization Approach*, M.I.T. Press, Cambridge, MA., 1985.
- [36] G. ZAMES AND B. A. FRANCIS, *Feedback, minimax sensitivity and optimal robustness*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 585–601.
- [37] K. ZHOU AND P. P. KHARGONEKAR, *A algebraic Riccati equation approach to \mathcal{H}_∞ optimization*, Systems Control Lett., 11 (1988), pp. 85–92.

RATE-PRESERVING DISCRETIZATION STRATEGIES FOR SEMI-INFINITE PROGRAMMING AND OPTIMAL CONTROL*

ELIJAH POLAK† AND LIMIN HE†

Abstract. Neither semi-infinite programming nor optimal control problems can be solved without discretization, i.e., decomposition of the original problems into an infinite sequence of finite-dimensional, finitely described optimization problems. Three sets of discretization refinement rules are presented: (i) for unconstrained semi-infinite minimax problems, (ii) for constrained semi-infinite problems, and (iii) for unconstrained optimal control problems. These rules are built into a master algorithm that calls certain linearly converging algorithms for finite-dimensional, finitely described optimization problems. The discretization refinement rules ensure that the sequences constructed by the overall scheme converge to a solution of the original problem linearly, with the estimated rate constant equal to the estimated rate constant of the algorithms used to solve the finite-dimensional, finitely described approximations. Hence the resulting scheme has the potential to be more efficient than fixed discretization, a fact supported by numerical results.

Key words. approximation theory, minimax, semi-infinite programming, optimal control, rate of convergence

AMS(MOS) subject classifications. 49K35, 49M39, 49J15

1. Introduction. The numerical solution of a semi-infinite optimization or optimal control problem always involves some form of discretization. There is considerable empirical evidence to suggest that the computationally most efficient approach is to increase discretization gradually. The heuristic explanation of the success of this approach is that far from a solution, coarse discretization does not appear to interfere with progress toward a solution, while resulting in considerable computational savings per iteration over fine discretization.

There are two basic, but not altogether disjoint, approaches to the construction of discretization refinements tests. The first is based on the concept of *diagonalization* (see, e.g., [4]–[6]), which can be used under minimal consistency conditions. Diagonalization decomposes the original optimization problem \mathbf{P} into an infinite sequence of *finite-dimensional, finitely described* optimization problems \mathbf{P}_q , $q = q_0, q_0 + 1, q_0 + 2, q_0 + 3, \dots$, which approximate \mathbf{P} more and more closely and which are therefore of increasing computational complexity¹. The problem \mathbf{P}_q is solved until a certain test is satisfied, and then the last iterate is used to initialize the solution of \mathbf{P}_{q+1} . For example, suppose that there is a family of continuous, negative-valued functions $\{\theta_q(\cdot)\}_{q=q_0}^\infty$ such that if \hat{x}_q is optimal for \mathbf{P}_q , then $\theta_q(\hat{x}_q) = 0$, and suppose that there is a continuous function $\theta(\cdot)$ such that (i) $\theta_q(x) \rightarrow \theta(x)$ as $q \rightarrow \infty$, uniformly in x (in a bounded set), and (ii) if \hat{x} is a solution to \mathbf{P} , then $\theta(\hat{x}) = 0$. Then an example of diagonalization scheme would consist of computing points x_q such that $\theta_q(x_q) \geq -1/q$. The main disadvantage of such a diagonalization approach is that all the convergence statements are in terms of the points x_q only (discarding the intermediate points constructed on the way to x_q); e.g., “all the accumulation points of the sequence

* Received by the editors October 23, 1989; accepted for publication (in revised form) March 4, 1991. This research was sponsored in part by National Science Foundation grant ECS-8713334, Air Force Office of Scientific Research contract AFOSR-86-0116, and State of California Microelectronics Innovation and Computer Research Opportunities (MICRO) Program grant 532410-19900.

† Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, California 94720.

¹ By this we mean that one iteration of a particular algorithm on problem \mathbf{P}_{q+1} is more costly than one iteration of the same algorithm on problem \mathbf{P}_q .

$\{x_q\}_{q=q_0}^\infty$ are stationary points for the problem \mathbf{P} .” The second approach is more subtle: it starts with a *conceptual* algorithm for solving \mathbf{P} (see [7], [10]) and uses progressively more precise numerical evaluations of the cost and constraint functions, as well as of the conceptual iteration map as an *implementation* of the conceptual algorithm. Implementation requires stronger consistency conditions than diagonalization; these, fortunately, are frequently satisfied in practice. In return, it yields the considerable advantage that all the convergence statements are in terms of the *entire* sequence constructed by the algorithm; e.g., “all the accumulation points of the entire sequence $\{x_i\}_{i=0}^\infty$ constructed by the implementation are stationary points for the problem \mathbf{P} .” In [7], [10] we find an abstract theory for the construction of *implementable* algorithms. As a particular application, we find in [7] an implementation of the method of steepest descent for solving continuous optimal control problems. Interesting recent examples of work dealing either with diagonalization or implementation of conceptual algorithms for optimal control and Banach space problems can be found in [5], [18], [19]. The results in [5], [18], [19] relate convergence rates for grid refinement schemes to the convergence rates of an associated gradient-method-related algorithm for some underlying limiting problem. In this sense, this work goes beyond the results in [7] and can be viewed as a precursor of the work presented in this paper.

The reason diagonalization and implementation schemes are not totally disjoint, is that when suitable tests of precision augmentation are used, diagonalization schemes also become implementation schemes. Such diagonalization/implementation schemes appear to be computationally more efficient than pure implementation schemes. The reason for this is suggested by the following example. Suppose that we wish to solve an unconstrained optimal control problem of the form $\min f(u)$ using the Armijo version of the steepest descent algorithm. A pure implementation of this algorithm may use unrelated numerical integration methods for computing approximations $\phi(u)$, $g(u)$ to $f(u)$, $\nabla f(u)$, and hence $g(u) \neq \nabla \phi(u)$. At low to medium precision of integration, this causes the implementable algorithm to slow down quite considerably, as compared to a diagonalization/implementation scheme that always ensures that $g(u) = \nabla \phi(u)$.

The discretization adjustment tests described in [7], [10] are very basic: they make no use of optimality functions or rate of convergence properties of the conceptual algorithm. In this paper, we present three diagonalization/implementation schemes in the form of master discretization algorithms that differ from those based on the theory in [7], [10] in two respects: (i) they use tests, based on optimality functions and a measure of the accuracy with which the discretized problems approximate the original problem, for increasing discretization, and (ii) unlike master algorithms based on the theory in [7], [10], under suitable assumptions, they preserve the estimates of convergence rate constants of the conceptual algorithms that they implement. The reason for their superiority is as follows. First-order algorithms are characterized by an estimate of the *rate constant* $\eta \in (0, 1)$, which appears in formulas of the form $e_{i+1} \leq \eta e_i$. The estimates of the rate constant of a first-order algorithm, applicable to discretized problems \mathbf{P}_q , depend on basic constants such as bounds on second-order derivatives and the Fritz John multiplier associated with the cost function. An examination of particular examples of discretized problems \mathbf{P}_q shows that the entire family $\{\mathbf{P}_q\}$ shares the *same* values of these constants that are inherited from the original problem \mathbf{P} (see, e.g., [5], [12]). Hence the estimate of the linear rate constant for a particular first-order algorithm is the *same* for *every* member of the family $\{\mathbf{P}_q\}$. Now suppose that a master discretization algorithm (say \mathbf{M}), which calls a particular first-order algorithm (say \mathbf{A}) as a subroutine, is linearly converging on \mathbf{P} , with the *same* rate constant that \mathbf{A} has

on the problems \mathbf{P}_q . Then, given an initial point x_0 , k iterations of \mathbf{M} on \mathbf{P} yield an endpoint $x_k^{\mathbf{M}}$ and discretization parameter q_k , while k iterations of \mathbf{A} on the problem $\mathbf{P}_{q_k}^{\mathbf{A}}$ yield an endpoint $x_k^{\mathbf{A}}$. Because of the same rate of convergence, we can expect that $x_k^{\mathbf{M}}$ and $x_k^{\mathbf{A}}$ are equally good approximations to a solution of \mathbf{P} . If we ignore the overhead imposed by the discretization tests, the total computing time used to produce $x_k^{\mathbf{M}}$ must be less because the early iterations of \mathbf{A} , as called by \mathbf{M} , face coarser discretizations than those encountered by \mathbf{A} in solving \mathbf{P}_{q_k} directly. Our experimental results indicate that the computing cost due to the discretization tests is not significant enough to affect our qualitative conclusion that the master discretization algorithm is more efficient than a fixed discretization scheme.

In § 2 we present a master discretization algorithm, to be used in conjunction with the Pshenichnyi–Pironneau–Polak (PPP) minimax algorithm [9], [12], [15] for solving unconstrained semi-infinite minimax problems. Its estimated rate-of-convergence constant is shown to be the same as that of the PPP algorithm, established in [9], [12]. In § 3 we present a master discretization algorithm, to be used in conjunction with the Polak–He unified steerable phase I–phase II method of feasible directions (USFD)² [13], for solving constrained semi-infinite optimization problems. Its estimated rate-of-convergence constant is shown to be the same as that of USFD. In § 4 we present a master discretization algorithm, to be used in conjunction with the Armijo gradient method [1], for solving unconstrained optimal control problems. Its estimated rate-of-convergence constant is shown to be the same as that of the Armijo method on composite function problems, established in [14].³ Finally, in § 5 we present some numerical results that support our qualitative conclusions as to the superiority of the new discretization schemes over fixed discretization approaches.

We use standard notation; thus $L_\infty^m[0, T]$ denotes the space of equivalence classes of essentially bounded, measurable functions from $[0, T]$ into \mathbb{R}^m , $L_2^m[0, T]$ denotes the space of equivalence classes of square integrable functions from $[0, T]$ into \mathbb{R}^m , and $\|\cdot\|$, $\langle \cdot, \cdot \rangle$ denote the Euclidean norm and scalar product, respectively, in \mathbb{R}^n . For $A \in \mathbb{R}^{m \times n}$, $\|A\| \triangleq \max_{\|x\|=1} \|Ax\|$, for $u, v \in L_2^m[0, T]$, $\|u\|_2^2 \triangleq \int_0^T \|u(t)\|^2 dt$, $\langle u, v \rangle_2 \triangleq \int_0^T \langle u(t), v(t) \rangle dt$, for $u \in L_\infty^m[0, T]$, $\|u\|_\infty \triangleq \text{ess sup}_{t \in [0, T]} \|u(t)\|$, and for $U \in L_\infty^{m \times n}[0, T]$, $\|U\|_\infty \triangleq \text{ess sup}_{t \in [0, T]} \|U(t)\|$.

2. Minimax problems. We begin with minimax problems of the form

$$(2.1a) \quad \text{MMP: } \min_{x \in \mathbb{R}^n} \psi(x),$$

where

$$(2.1b) \quad \psi(x) \triangleq \max_{j \in \mathbf{I}} \max_{y_j \in \mathbf{Y}_j} \phi^j(x, y_j),$$

where $\mathbf{I} \triangleq \{1, 2, \dots, l\}$, $\phi^j: \mathbb{R}^n \times \mathbb{R}^{p_j} \rightarrow \mathbb{R}$ and \mathbf{Y}_j is a compact subset of \mathbb{R}^{p_j} . We will assume that the functions $\phi^j(\cdot, \cdot)$ and their gradients $\nabla_x \phi^j(\cdot, \cdot)$ are Lipschitz continuous. Since the exact calculation of the global maxima of $\phi^j(\cdot, \cdot)$ over the compact

² This algorithm is the only phase I–phase II method of feasible directions that we were able to implement in such a way that once a feasible point for a problem P_{q_0} was found, terminating phase I on P_{q_0} , the algorithm remained in phase II for this and all the following problems P_q .

³ Discrete optimal control problems have cost functions of the form $f(u) = g(x(N, u, x_0))$, and the Hessian of $f(\cdot)$ is positive-semi-definite, at best. Hence “standard” rate of convergence theory (see [8], [10]) leads to the conclusion that the Armijo method converges on these problems sublinearly. The results for linear dynamics, in [14], to be extended in this paper, show that this is not so.

set Y_j is not a numerically implementable operation, numerical methods for solving problem (2.1a) must discretize the compact sets Y_j . Hence we introduce a family of approximating problems, parametrized by the discretization parameter $q \in \mathbb{N}$:

$$(2.2a) \quad \mathbf{MMP}_q : \min_{x \in \mathbb{R}^n} \psi_q(x),$$

where

$$(2.2b) \quad \psi_q(x) \triangleq \max_{j \in I} \max_{y_j \in Y_j} \phi_q^j(x, y_j),$$

and the functions $\phi_q^j(\cdot, \cdot)$ are constructed by linear interpolation of the $\phi^j(\cdot, \cdot)$ over a “triangulated” (uniform) grid in the sets Y_j . Thus, for example, when $Y_j = [0, 1] \subset \mathbb{R}$, we divide this interval into q subintervals, and then we define $\phi_q^j(x, y)$ to be linear on each interval, so that

$$\begin{aligned} \phi_q^j(x, y) &= \lambda \phi^j(x, i/q) + (1 - \lambda) \phi^j(x, (i+1)/q) \\ \text{for } y &= \lambda i/q + (1 - \lambda)(i+1)/q, \quad \lambda \in [0, 1], \quad i = 0, 1, 2, \dots, q-1 \end{aligned}$$

When $Y_j = [0, 1] \times [0, 1] \subset \mathbb{R}^2$, we first break up the plane into small squares, with sides of length $1/q$, and each square is then divided into two triangles, using parallel diagonals. The function $\phi_q^j(x, y)$ is then defined as a continuous linear interpolation of $\phi^j(x, y)$ on this triangulated grid. We note that when defined in this manner, the evaluation of $\psi_q(x)$ is a finite process.

The master adaptive discretization algorithm that we will shortly introduce, calls the PPP minimax algorithm [9], [11], [15] as a subroutine. This algorithm computes search directions by evaluating an optimality function. Hence, proceeding as in [11], we define $\theta(x)$, $\theta_q(x)$ to be the *optimality functions* for problems **MMP** and problem **MMP**_q, respectively, as follows:

$$(2.3a) \quad \theta(x) \triangleq \min_{h \in \mathbb{R}^n} \max_{j \in I} \max_{y_j \in Y_j} \{ \phi^j(x, y_j) + \langle \nabla_x \phi^j(x, y_j), h \rangle + \tfrac{1}{2} \|h\|^2 - \psi(x) \},$$

$$(2.3b) \quad \theta_q(x) \triangleq \min_{h \in \mathbb{R}^n} \max_{j \in I} \max_{y_j \in Y_j} \{ \phi_q^j(x, y_j) + \langle \nabla_x \phi_q^j(x, y_j), h \rangle + \tfrac{1}{2} \|h\|^2 - \psi_q(x) \}.$$

When the functions $\phi_q^j(\cdot, \cdot)$ are defined by linear interpolation on a triangulated grid, (2.3b) is an ordinary quadratic programming problem that can be solved finitely using standard quadratic programming subroutines.

Let $h, h_q : \mathbb{R}^n \rightarrow \mathbb{R}$ be *search direction functions* defined by

$$(2.3c) \quad h(x) \triangleq \arg \min_{h \in \mathbb{R}^n} \max_{j \in I} \max_{y_j \in Y_j} \{ \phi^j(x, y_j) + \langle \nabla_x \phi^j(x, y_j), h \rangle + \tfrac{1}{2} \|h\|^2 - \psi(x) \},$$

$$(2.3d) \quad h_q(x) \triangleq \arg \min_{h \in \mathbb{R}^n} \max_{j \in I} \max_{y_j \in Y_j} \{ \phi_q^j(x, y_j) + \langle \nabla_x \phi_q^j(x, y_j), h \rangle + \tfrac{1}{2} \|h\|^2 - \psi_q(x) \}.$$

The PPP minimax algorithms for solving **MMP** and **MMP**_q use one of the above search direction functions (as appropriate) and an Armijo-type stepsize rule, which requires two parameters $\alpha, \beta \in (0, 1)$. Thus the *conceptual* PPP algorithm for solving **MMP**, described in [11], constructs iterates according to the rule

$$(2.3e) \quad x_{i+1} = x_i + \lambda_i h(x_i),$$

where

$$(2.3f) \quad \lambda_i = \max \{ \beta^k \mid k \in \mathbb{N}, \psi(x_i + \beta^k h(x_i)) - \psi(x_i) \leq \alpha \beta^k \theta(x_i) \},$$

while the *implementable* PPP algorithm for solving \mathbf{MMP}_q , described in [9], [11], [15], constructs iterates according to (2.3e), (2.3f), with $h(\cdot)$, $\psi(\cdot)$, and $\theta(\cdot)$ replaced by $h_q(\cdot)$, $\psi_q(\cdot)$, and $\theta_q(\cdot)$, respectively.

No matter how the approximating functions $\phi_q^j(\cdot, \cdot)$ are constructed, we will require that the functions $\phi_q^j(\cdot, \cdot)$, together with the functions $\phi^j(\cdot, \cdot)$, satisfy the following assumption.⁴

ASSUMPTION 2.1. (i) There exist constants $0 < K < \infty$ and $\tau > 0$ such that for all $x \in \mathbb{R}^n$ and all $q \in \mathbb{N}$,

$$(2.4a) \quad |\psi(x) - \psi_q(x)| \leq K/q^\tau.$$

(ii) For any $x \in \mathbb{R}^n$, $y_j \in \mathbf{Y}_j$, and $j \in \mathbf{I}$,

$$(2.4b) \quad \lim_{\substack{x' \rightarrow x \\ q \rightarrow \infty}} \phi_q^j(x', y_j) = \phi^j(x, y_j), \quad \lim_{\substack{x' \rightarrow x \\ q \rightarrow \infty}} \nabla_x \phi_q^j(x', y_j) = \nabla_x \phi^j(x, y_j).$$

LEMMA 2.1. (i) For any $x \in \mathbb{R}^n$,

$$(2.5a) \quad \theta(x) = -\min \{ \xi^0 + \frac{1}{2} \|\xi\|^2 \mid (\xi^0, \xi)^T \in G(x) \},$$

$$(2.5b) \quad \theta_q(x) = -\min \{ \xi^0 + \frac{1}{2} \|\xi\|^2 \mid (\xi^0, \xi)^T \in G_q(x) \},$$

where

$$(2.5c) \quad G(x) \triangleq \text{co} \left\{ \bigcup_{j \in \mathbf{I}} \bigcup_{y_j \in \mathbf{Y}_j} \begin{pmatrix} \psi(x) - \phi^j(x, y_j) \\ \nabla_x \phi^j(x, y_j) \end{pmatrix} \right\},$$

$$(2.5d) \quad G_q(x) \triangleq \text{co} \left\{ \bigcup_{j \in \mathbf{I}} \bigcup_{y_j \in \mathbf{Y}_j} \begin{pmatrix} \psi_q(x) - \phi_q^j(x, y_j) \\ \nabla_x \phi_q^j(x, y_j) \end{pmatrix} \right\}.$$

(ii) For any $x \in \mathbb{R}^n$, $\theta(x) = 0$ if and only if $0 \in \partial \psi(x)$,⁵ and $\theta_q(x) = 0$ if and only if $0 \in \partial \psi_q(x)$; i.e., the zeros of these functions are the stationary points of the corresponding problems.

(iii) Suppose that Assumption 2.1 holds. Then, for any $x \in \mathbb{R}^n$,

$$(2.6) \quad \lim_{\substack{x' \rightarrow x \\ q \rightarrow \infty}} \theta_q(x') = \theta(x).$$

Proof. Both (i) and (ii) are established in [11].

(iii) Since the sets \mathbf{Y}_j are compact, it follows from Assumption 2.1(ii) and (2.5c), (2.5d) that

$$(2.7) \quad \lim_{\substack{x' \rightarrow x \\ q \rightarrow \infty}} G_q(x') = G(x).$$

Hence (2.6) follows from (2.7) and the definitions (2.5a), (2.5b). \square

We can now state a master adaptive discretization algorithm that calls the PPP minimax algorithm as a subroutine, for solving the problem \mathbf{MMP} .

⁴ This assumption is satisfied, with $\tau = 1$, by the two examples we gave using a *uniform* discretization grid, when the functions $\phi^j(x, \cdot)$ are at least Lipschitz continuous. When the functions $\nabla_x \phi^j(x, \cdot)$ are Lipschitz continuous, then our assumption is satisfied with $\tau = 2$.

⁵ We denote the Clarke-generalized gradient [3] of a locally Lipschitz function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ at $x \in \mathbb{R}^n$ by $\partial f(x)$.

Adaptive Discretization Algorithm 2.1 (for MMP):

Data: $x_0 \in \mathbb{R}^n$, $q_{-1} \in \mathbb{N}$, $\alpha \in (0, 1)$, $\beta \in (0, 1)$, $D > 0$, and $\sigma > 1$.

Step 0: Set $i = 0$.

Step 1: Compute $q_i \in \mathbb{N}$, $\theta_{q_i}(x_i)$, and, simultaneously, $h_{q_i}(x_i)$ such that $q_i \geq q_{i-1}$ and

$$(2.8) \quad D/q_i^\sigma \leq [-\theta_{q_i}(x_i)]^\sigma.$$

Step 2: Set $\theta_i = \theta_{q_i}(x_i)$, $h_i = h_{q_i}(x_i)$ and compute the stepsize λ_i :

$$(2.9) \quad \lambda_i = \max \{\beta^k \mid k \in \mathbb{N}, \psi_{q_i}(x_i + \beta^k h_i) - \psi_{q_i}(x_i) \leq \alpha \beta^k \theta_i\}.$$

Step 3: Set $x_{i+1} = x_i + \lambda_i h_i$, replace i by $i + 1$, and go to Step 1.

Remark 2.1. (a) The discretization rule (2.8) is chosen to ensure both convergence and preservation of rate of convergence. As can be seen from [7], mere convergence can be ensured by other tests as well.

(b) It follows from Lemma 2.1(iii) that whenever $\theta(x_i) \neq 0$, Step 1 of Algorithm 2.1 yields a finite discretization parameter q_i . For simplicity, in the rest of this section we assume that Algorithm 2.1 does not produce an iterate x_i such that $\theta(x_i) = 0$, so that the resulting value of q_i in Step 1 is finite.

LEMMA 2.2. *Suppose that $\psi(\cdot)$ is bounded from below and that the sequence of iterates $\{x_i\}_{i=0}^\infty$ and corresponding sequence of discretization parameters $\{q_i\}_{i=0}^\infty$ were constructed by Algorithm 2.1. Then $q_i \rightarrow \infty$ as $i \rightarrow \infty$.*

Proof. For $i \in \mathbb{N}$, let q_i , θ_i , h_i , and λ_i be defined as in Algorithm 2.1, and suppose that $q_i \rightarrow \infty$ as $i \rightarrow \infty$ does not hold. Then, since $\{q_i\}_{i=0}^\infty$ is a nondecreasing sequence of integers, it follows that there exist i_0 , $\hat{q} \in \mathbb{N}$, such that for all $i \geq i_0$, $q_i = \hat{q}$, and hence, in conjunction with (2.8), that there exists an $\varepsilon > 0$, such that $\theta_i \leq -\varepsilon$ for all $i \geq i_0$. Making use of (2.9) and the assumption that $\psi(\cdot)$ is bounded from below, we conclude that $\psi_{\hat{q}}(\cdot)$ is also bounded from below. Hence we obtain that

$$-\infty < \sum_{i=i_0}^{\infty} [\psi_{\hat{q}}(x_{i+1}) - \psi_{\hat{q}}(x_i)] \leq \sum_{i=i_0}^{\infty} \alpha \lambda_i \theta_i.$$

Referring to (2.5b), we see that if

$$(\xi_i^0, \xi_i) = \arg \min \{ \xi_i^0 + \frac{1}{2} \|\xi\|^2 \mid (\xi_i^0, \xi) \in G_{q_i}(x_i) \},$$

then $h_i = -\xi_i$. Hence $\|h_i\|^2 \leq -2\theta_i$ and $-\theta_i \geq \varepsilon$ for $i \geq i_0$, we deduce that $\|h_i\|^2 \leq 2\theta_i^2/\varepsilon$ for all $i \geq i_0$. Hence, for any $j > i \geq i_0$,

$$(2.10) \quad \|x_j - x_i\| \leq \sum_{k=i}^{j-1} \|x_{k+1} - x_k\| \leq \sum_{k=i}^{j-1} \lambda_k \|h_k\| \leq \sum_{k=i}^{\infty} (2/\varepsilon)^{1/2} \lambda_k (-\theta_k).$$

Therefore $\{x_i\}_{i=0}^\infty$ is a Cauchy sequence in \mathbb{R}^n , and hence it follows from Theorem 5.2(b) and Corollary 5.1 in [11] (which show that any accumulation point x^* of $\{x_i\}_{i=0}^\infty$, constructed by the PPP algorithm, satisfies $\theta_{\hat{q}}(x^*) = 0$) that $\theta_{\hat{q}}(x_i) \rightarrow 0$, contradicting the construction in (2.8). \square

THEOREM 2.1. *Suppose that Assumption 2.1 holds, that $\psi(\cdot)$ is bounded from below, that the second derivatives $\partial^2 \phi_q^j(x, y_j)/\partial x^2$ exist for all $q \in \mathbb{N}$, and that there exists an $M \in [1, \infty)$ such that for all $x \in \mathbb{R}^n$, $z \in \mathbb{R}^n$, $j \in \mathbb{I}$, and $q \in \mathbb{N}$,*

$$(2.11) \quad \left\langle z, \frac{\partial^2 \phi_q^j(x, y_j)}{\partial x^2} z \right\rangle \leq M \|z\|^2, \quad \forall y_j \in \mathbf{Y}_j.$$

Then any accumulation point \hat{x} of the sequence of iterates $\{x_i\}_{i=0}^\infty$, generated by Algorithm 2.1, satisfies $\theta(\hat{x}) = 0$.

Proof. First, we obtain a bound on the decrease in $\psi(\cdot)$ at the i th iteration. Using (2.11), we obtain that

$$\begin{aligned}
 \psi_{q_i}(x_i + \lambda h_i) - \psi_{q_i}(x_i) &= \max_{j \in \mathbf{I}} \max_{y_j \in \mathbf{Y}_j} \{ \phi_{q_i}^j(x_i + \lambda h_i, y_j) - \psi_{q_i}(x_i) \} \\
 (2.12) \quad &\leq \max_{j \in \mathbf{I}} \max_{y_j \in \mathbf{Y}_j} \{ \phi_{q_i}^j(x_i, y_j) - \psi_{q_i}(x_i) + \langle \nabla_x \phi_{q_i}^j(x_i, y_j), \lambda h_i \rangle \\
 &\quad + \tfrac{1}{2} M \| \lambda h_i \|^2 \}.
 \end{aligned}$$

In view of (2.3b) and the fact that $\phi_{q_i}^j(x_i, y_j) - \psi_{q_i}(x_i) \leq 0$ and that $M \geq 1$, we find that for all $\lambda \in [0, 1/M]$,

$$(2.13) \quad \psi_{q_i}(x_i + \lambda h_i) - \psi_{q_i}(x_i) \leq \lambda \theta_{q_i}(x_i).$$

Therefore (2.9) is satisfied with $\lambda_i \geq \beta/M$, and thus

$$(2.14) \quad \psi_{q_i}(x_{i+1}) - \psi_{q_i}(x_i) \leq \alpha \lambda_i \theta_i \leq \alpha \beta \theta_i / M.$$

Hence it follows from Assumption 2.1(i) that

$$(2.15) \quad \psi(x_{i+1}) - \psi(x_i) \leq \alpha \beta \theta_i / M + 2K/q_i^\tau.$$

Next, since $\theta_i \leq -D^{1/\sigma}/(q_i^\tau)^{1/\sigma}$, we have that

$$(2.16) \quad \psi(x_{i+1}) - \psi(x_i) \leq -\frac{\alpha \beta D^{1/\sigma}}{M(q_i^\tau)^{1/\sigma}} \left[1 - \frac{2KM}{\alpha \beta D^{1/\sigma}(q_i^\tau)^{(\alpha-1)/\sigma}} \right].$$

Since $\sigma > 1$ and since by Lemma 2.2, $q_i \rightarrow \infty$ as $i \rightarrow \infty$, there exists an $i_0 \in \mathbb{N}$ such that for all $i \geq i_0$,

$$(2.17) \quad \psi(x_{i+1}) - \psi(x_i) \leq -\frac{\alpha \beta D^{1/\sigma}}{2M(q_i^\tau)^{1/\sigma}}.$$

Hence,

$$(2.18) \quad \psi(x_{i+1}) - \psi(x_{i_0}) \leq -\sum_{k=i_0}^i \frac{\alpha \beta D^{1/\sigma}}{2m(q_k^\tau)^{1/\sigma}}.$$

Because $\psi(\cdot)$ is bounded from below, the left-hand side of the above inequality is bounded from below, which leads to the conclusion that $\sum_{k=0}^{\infty} 1/(q_k^\tau)^{1/\sigma} < \infty$. Consequently, $\sum_{k=0}^{\infty} 1/(q_k^\tau) < \infty$. Next, returning to (2.15), we conclude that

$$(2.19) \quad \psi(x_{i+1}) - \psi(x_0) \leq -\sum_{k=0}^i \alpha \beta (-\theta_k) / M + \sum_{k=0}^i 2K/q_k^\tau.$$

Since both $\psi(x_{i+1}) - \psi(x_0)$ and $\sum_{k=0}^i 2K/q_k^\tau$ are bounded, it follows that $\sum_{k=0}^{\infty} (-\theta_k) < \infty$. The desired result now follows from Lemma 2.1(iii). \square

THEOREM 2.2. *Suppose that Assumption 2.1 holds, that the second derivatives $\partial^2 \phi^j(x, y_j)/\partial x^2$, $\partial^2 \phi_q^j(x, y_j)/\partial x^2$ exist for all $q \in \mathbb{N}$, and that there exist constants $0 < m < 1 < M < \infty$, such that for all $x \in \mathbb{R}^n$, $z \in \mathbb{R}^n$, and $j \in \mathbf{I}$,*

$$(2.20a) \quad m \|z\|^2 \leq \left\langle z, \frac{\partial^2 \phi_q^j(x, y_j)}{\partial x^2} z \right\rangle \leq M \|z\|^2 \quad \text{for all } y_j \in \mathbf{Y}_j, \quad q \in \mathbb{N},$$

$$(2.20b) \quad m \|z\|^2 \leq \left\langle z, \frac{\partial^2 \phi^j(x, y_j)}{\partial x^2} z \right\rangle \leq M \|z\|^2 \quad \text{for all } y_j \in \mathbf{Y}_j.$$

Then any sequence $\{x_i\}_{i=0}^\infty$, generated by Algorithm 2.1, converges to the unique solution \hat{x} of problem **MMP**, and

$$(2.21) \quad \lim_{i \rightarrow \infty} \frac{\psi(x_{i+1}) - \psi(\hat{x})}{\psi(x_i) - \psi(\hat{x})} \leq 1 - \alpha\beta m / M.$$

Proof. It follows from (2.20a), (2.20b) that the functions $\psi_q(\cdot)$ $q \in \mathbb{N}$, $\psi(\cdot)$ are strongly convex, and hence that they have unique minimizers. For any $q \in \mathbb{N}$, let \hat{x}_q be the unique solution of the problem **MMP** _{q} . First, we deduce from Assumption 2.1(i) that

$$(2.22) \quad |\psi_q(\hat{x}_q) - \psi(\hat{x})| \leq K / q^\tau.$$

Next, referring to [12], [14], we see that for all $x \in \mathbb{R}^n$ and $q \in \mathbb{N}$,

$$(2.23) \quad m[\psi_q(x) - \psi_q(\hat{x}_q)] \leq -\theta_q(x) \leq M[\psi_q(x) - \psi_q(\hat{x}_q)].$$

Combining (2.14) and (2.23), we get

$$(2.24) \quad \psi_{q_i}(x_{i+1}) - \psi_{q_i}(x_i) \leq \frac{\alpha\beta m}{M} [\psi_{q_i}(\hat{x}_{q_i}) - \psi_{q_i}(x_i)].$$

Adding $\psi_{q_i}(\hat{x}_{q_i})$ to both sides in (2.24) and rearranging terms, we obtain that

$$(2.25) \quad \psi_{q_i}(x_{i+1}) - \psi_{q_i}(\hat{x}_{q_i}) \leq [1 - \alpha\beta m / M][\psi_{q_i}(x_i) - \psi_{q_i}(\hat{x}_{q_i})].$$

It now follows from (2.22) and Assumption 2.1(i) that

$$(2.26) \quad \psi(x_{i+1}) - \psi(\hat{x}) \leq [1 - \alpha\beta m / M][\psi(x_i) - \psi(\hat{x})] + 4K / q_i^\tau.$$

Next, making use of Assumption 2.1(i), (2.22), (2.23), and the fact that $-\theta_i \geq D^{1/\sigma} / (q_i^\tau)^{1/\sigma}$, we obtain that

$$(2.27) \quad \begin{aligned} M[\psi(x_i) - \psi(\hat{x})] &\geq M[\psi_{q_i}(x_i) - \psi(\hat{x}_{q_i})] - 2MK / q_i^\tau \\ &\geq -\theta_i - 2MK / q_i^\tau \\ &\geq D^{1/\sigma} / (q_i^\tau)^{1/\sigma} - 2MK / q_i^\tau. \end{aligned}$$

Since $\sigma > 1$ and since by Lemma 2.2, $q_i \rightarrow \infty$ as $i \rightarrow \infty$, we conclude that there exists i_0 such that for all $i \geq i_0$,

$$(2.28) \quad M[\psi(x_i) - \psi(\hat{x})] \geq D^{1/\sigma} / [2(q_i^\tau)^{1/\sigma}].$$

It now follows from (2.26) and (2.28) that

$$(2.29) \quad \psi(x_{i+1}) - \psi(\hat{x}) \leq \left[1 - \frac{\alpha\beta m}{M} + \frac{8KM}{D^{1/\sigma}(q_i^\tau)^{(\sigma-1)/\sigma}} \right] [\psi(x_i) - \psi(\hat{x})] \quad \text{for } i \geq i_0.$$

Therefore (2.21) follows from (2.29) and the fact that by Theorem 2.1, $q_i \rightarrow \infty$ as $i \rightarrow \infty$. Since $\psi(x_i) \rightarrow \psi(\hat{x})$ and \hat{x} is the unique minimizer of $\psi(\cdot)$, $\{x_i\}_{i=0}^\infty$ must converge to \hat{x} . \square

For comparison, referring to [12], we find the following result for the *conceptual* PPP minimax algorithm.

THEOREM 2.3.⁶ Suppose that the second derivatives $\partial^2 \phi^j(x, y_j) / \partial x^2$ exist and that there exist constants $0 < m < 1 < M < \infty$, such that for all $x \in \mathbb{R}^n$, $z \in \mathbb{R}^n$, and $j \in \mathbf{1}$, (2.20b) is satisfied. Then any sequence $\{x_i\}_{i=0}^\infty$, generated by the PPP minimax algorithm defined by (2.3e), (2.3f), converges to the unique solution \hat{x} of problem **MMP**, and (2.21) holds.

⁶ It should be obvious that, under analogous assumptions, the conclusions of Theorem 2.3 remain valid for the implementable PPP minimax algorithm, which can be used to solve **MMP** _{q} .

We thus see that our Adaptive Discretization Algorithm 2.1 converges with exactly the same linear rate constant as the PPP minimax algorithm, whether applied to **MMP** or to any **MMP**_q. This leads us to the following observation as to the relative efficiency of our Adaptive Discretization Algorithm 2.1, under the assumptions in Theorem 2.3. Suppose that we are given an initial point x_0 (sufficiently close to \hat{x} , the solution of **MMP**), and that we perform k iterations, ending with a point x_k and discretization parameter q_k . If, instead, we perform k iterations on the problem **MMP**_{q_k}, ending at a point x'_k , we cannot expect to have $\psi(x'_k) < \psi(x_k)$. However, the total computing time used on solving **MMP**_{q_k} must be longer because the early iterations of Algorithm 2.1 use a coarser discretization. Hence Algorithm 2.1 has the potential for being more efficient than a fixed discretization scheme.

3. Constrained semi-infinite optimization problems. We will now consider constrained semi-infinite optimization problems of the form

$$(3.1a) \quad \text{CSP: } \min \{ \psi^0(x) \mid \psi^j(x) \leq 0, j \in \mathbf{I}, x \in \mathbb{R}^n \},$$

where $\mathbf{I} \triangleq \{1, 2, \dots, l\}$, and, with $\mathbf{L} \triangleq \{0, 1, 2, \dots, l\}$,

$$(3.1b) \quad \psi^j(x) = \max_{y_j \in \mathbf{Y}_j} \phi^j(x, y_j), \quad \forall j \in \mathbf{L},$$

where $\phi^j: \mathbb{R}^n \times \mathbb{R}^{p_j} \rightarrow \mathbb{R}$ and \mathbf{Y}_j is a compact set in \mathbb{R}^{p_j} . We will assume that the functions $\phi^j(\cdot, \cdot)$ and their gradients $\nabla_x \phi^j(\cdot, \cdot)$ are Lipschitz continuous.

Using the interpolation techniques mentioned in the preceding section, we can construct a family of approximating problems, parametrized by the discretization parameter $q \in \mathbb{N}$:

$$(3.2a) \quad \text{CSP}_q: \min \{ \psi_q^0(x) \mid \psi_q^j(x) \leq 0, j \in \mathbf{I}, x \in \mathbb{R}^n \},$$

where

$$(3.2b) \quad \psi_q^j(x) = \max_{y_j \in \mathbf{Y}_j} \phi_q^j(x, y_j), \quad \forall j \in \mathbf{L}.$$

In [13] we find a unified steerable phase I-phase II method of feasible directions, which uses a *steering parameter* $\gamma > 0$. We will refer to this algorithm as the USFD algorithm. The steering parameter controls the speed with which infeasible iterates approach the feasible set. When this parameter is greater than a certain value, the algorithm in [13] constructs a feasible point in a finite number of iterations. We will call this algorithm as a subroutine from the master adaptive discretization algorithm that we will describe shortly. For the original problem **CSP**, the algorithm in [13] requires the following functions:

$$(3.3a) \quad \psi(x) \triangleq \max_{j \in \mathbf{I}} \psi^j(x),$$

$$(3.3b) \quad \psi_+(x) \triangleq \max \{0, \psi(x)\},$$

$$(3.3c) \quad F_z(x) \triangleq \max \{ \psi^0(x) - \psi^0(z) - \gamma \psi_+(z), \psi(x) - \psi_+(z) \},$$

$$(3.3d) \quad \theta(x) \triangleq \min_{h \in \mathbb{R}^n} \max \left\{ \max_{y_0 \in \mathbf{Y}_0} \{ \phi^0(x, y_0) + \langle \nabla_x \phi^0(x, y_0), h \rangle + \tfrac{1}{2} \|h\|^2 - \psi^0(x) - \gamma \psi_+(x) \}, \right. \\ \left. \cdot \max_{j \in \mathbf{I}} \max_{y_j \in \mathbf{Y}_j} \{ \phi^j(x, y_j) + \langle \nabla_x \phi^j(x, y_j), h \rangle + \tfrac{1}{2} \|h\|^2 - \psi_+(x) \} \right\},$$

$$(3.3e) \quad h(x) \triangleq \arg \min_{h \in \mathbb{R}^n} \max \left\{ \max_{y_0 \in \mathbf{Y}_0} \{ \phi^0(x, y_0) + \langle \nabla_x \phi^0(x, y_0), h \rangle + \frac{1}{2} \|h\|^2 - \psi^0(x) - \gamma \psi_+(x) \} \right. \\ \left. \cdot \max_{j \in \mathbf{I}} \max_{y_j \in \mathbf{Y}_j} \{ \phi^j(x, y_j) + \langle \nabla_x \phi^j(x, y_j), h \rangle + \frac{1}{2} \|h\|^2 - \psi_+(x) \} \right\}.$$

For the problems \mathbf{CSP}_q , the algorithm in [13] uses corresponding quantities $\psi_q(\cdot)$, $\psi_{q+}(\cdot)$, $F_{z,q}(x)$, $\theta_q(x)$, and $h_q(x)$, resulting from the replacement of $\phi^j(\cdot, \cdot)$, $\nabla_x \phi^j(\cdot, \cdot)$, $\psi^j(\cdot)$, $\psi(\cdot)$, and $\psi_+(\cdot)$ by $\phi_q^j(\cdot, \cdot)$, $\nabla_x \phi_q^j(\cdot, \cdot)$, $\psi_q^j(\cdot)$, $\psi_q(\cdot)$, and $\psi_{q+}(x)$ in (3.3a)–(3.3e), respectively.

In addition to the steering parameter γ , the USFD algorithm in [13] uses two Armijo stepsize parameters: $\alpha, \beta \in (0, 1)$. In solving \mathbf{CSP} , this algorithm constructs iterates according to the rule

$$(3.3f) \quad x_{i+1} = x_i + \lambda_i h(x_i),$$

where

$$(3.3g) \quad \lambda_i = \max \{ \beta^k \mid k \in \mathbb{N}, F_{x_i}(x_i + \beta^k h(x_i)) - F_{x_i}(x_i) \leq \alpha \beta^k \theta(x_i) \}.$$

For solving \mathbf{CSP}_q , we make the obvious substitutions in the above rule.

We will assume that the following relationship between the functions defining the problems \mathbf{CSP}_q and problem \mathbf{CSP} .

Assumption 3.1. (i) There exist constants $0 < k < \infty$ and $\tau > 0$ such that for all $x \in \mathbb{R}^n$ and all $q \in \mathbb{N}$,

$$(3.4a) \quad |\psi^0(x) - \psi_q^0(x)| \leq K/q^\tau,$$

$$(3.4b) \quad |\psi(x) - \psi_q(x)| \leq K/q^\tau.$$

(ii) For any $x \in \mathbb{R}^n$, $y_j \in \mathbf{Y}_j$, and $j \in \mathbf{L}$,

$$(3.4c) \quad \lim_{\substack{x' \rightarrow x \\ q \rightarrow \infty}} \phi_q^j(x', y_j) = \phi^j(x, y_j), \quad \lim_{\substack{x' \rightarrow x \\ q \rightarrow \infty}} \nabla_x \phi_q^j(x', y_j) = \nabla_x \phi^j(x, y_j).$$

LEMMA 3.1. (i) For any $x \in \mathbb{R}^n$,

$$(3.5a) \quad \theta(x) = -\min \{ \xi^0 + \frac{1}{2} \|\xi\|^2 \mid (\xi^0, \xi)^T \in G(x) \},$$

$$(3.5b) \quad \theta_q(x) = -\min \{ \xi^0 + \frac{1}{2} \|\xi\|^2 \mid (\xi^0, \xi)^T \in G_q(x) \},$$

where

$$(3.5c) \quad G(x) \triangleq \text{co} \left\{ \bigcup_{y_0 \in \mathbf{Y}_0} \begin{pmatrix} \psi^0(x) - \phi^0(x, y_0) + \gamma \psi_+(x) \\ \nabla_x \phi^0(x, y_0) \end{pmatrix}, \right. \\ \left. \cdot \bigcup_{j \in \mathbf{I}} \bigcup_{y_j \in \mathbf{Y}_j} \begin{pmatrix} \psi(x) - \phi^j(x, y_j) \\ \nabla_x \phi^j(x, y_j) \end{pmatrix} \right\},$$

$$(3.5d) \quad G_q(x) \triangleq \text{co} \left\{ \bigcup_{y_0 \in \mathbf{Y}_0} \begin{pmatrix} \psi_q^0(x) - \phi_q^0(x, y_0) + \gamma \psi_{q+}(x) \\ \nabla_x \phi_q^0(x, y_0) \end{pmatrix}, \right. \\ \left. \cdot \bigcup_{j \in \mathbf{I}} \bigcup_{y_j \in \mathbf{Y}_j} \begin{pmatrix} \psi_q(x) - \phi_q^j(x, y_j) \\ \nabla_x \phi_q^j(x, y_j) \end{pmatrix} \right\}.$$

(ii) For any $x \in \mathbb{R}^n$, $\theta(x) = 0$ if and only if either $\psi(x) \leq 0$ and $0 \in \partial F_x(x)$ (i.e., x satisfies the first-order optimality condition for problem (3.1a)), or $\psi(x) > 0$ and $0 \in \partial \psi(x)$, (i.e., x satisfies the first-order optimality condition for the problem $\min_{x \in \mathbb{R}^n} \psi(x)$).

(iii) Suppose that Assumption 3.1 holds. Then, for any $x \in \mathbb{R}^n$,

$$(3.6) \quad \lim_{\substack{x' \rightarrow x \\ q \rightarrow \infty}} \theta_q(x') = \theta(x).$$

Proof. (i) Relations (3.5a), (3.5b) were established in [13] using von Neuman's minimax theorem.

(ii) This part can be deduced from Propositions 5.4 and 5.5 in [11].

(iii) Since \mathbf{Y}_j are compact sets, it follows from Assumption 3.1(ii) and (3.5c), (3.5d) that

$$(3.7) \quad \lim_{\substack{x' \rightarrow x \\ q \rightarrow \infty}} G_q(x') = G(x).$$

Hence (3.6) follows from (3.7) and (3.5a), (3.5b). \square

We are now ready to state an adaptive discretization scheme, based on the USFD algorithm in [13], for solving the problem CSP.

Adaptive Discretization Algorithm 3.1 (for CSP):

Data: $x_0 \in \mathbb{R}^n$, $q_{-1} \in \mathbb{N}$, $\alpha \in (0, 1)$, $\beta \in (0, 1)$, $\gamma > 0$, $D > 0$, and $\sigma > 1$.

Step 0: Set $i = 0$.

Step 1: Compute $q_i \in \mathbb{N}$, $\theta_{q_i}(x_i)$, and, simultaneously, $h_{q_i}(x_i)$, such that $q_i \geq q_{i-1}$ and

$$(3.8) \quad D/q_i^\gamma \leq [-\theta_{q_i}(x_i)]^\sigma$$

Step 2: Set $\theta_i = \theta_{q_i}(x_i)$, $h_i = h_{q_i}(x_i)$, and compute the stepsize λ_i :

$$(3.9) \quad \lambda_i = \max \{\beta^k \mid k \in \mathbb{N}, F_{x_i, q_i}(x_i + \beta^k h_i) - F_{x_i, q_i}(x_i) \leq \alpha \beta^k \theta_i\}.$$

Step 3: Set $x_{i+1} = x_i + \lambda_i h_i$, replace i by $i + 1$, and go to Step 1.

Remark 3.1. It follows from Lemma 3.1(ii)–(iii) that whenever $\theta(x_i) \neq 0$, Step 1 of Algorithm 3.1 yields a finite discretization parameter q_i . For simplicity, in the rest of this section we assume that Algorithm 3.1 does not produce an iterate x_i such that $\theta(x_i) = 0$, so that the resulting value of q_i in Step 1 is finite.

LEMMA 3.2. Suppose that $\psi^0(\cdot)$ is bounded from below and that the sequence of iterates $\{x_i\}_{i=0}^\infty$ and the corresponding sequence of discretization parameters $\{q_i\}_{i=0}^\infty$ were constructed by Algorithm 3.1. Then $q_i \rightarrow \infty$ as $i \rightarrow \infty$.

Proof. Suppose that $q_i \rightarrow \infty$ as $i \rightarrow \infty$ does not hold. Then, since $\{q_i\}_{i=0}^\infty$ is a nondecreasing sequence of integers, it follows that there exists an i_0 , $\hat{q} \in \mathbb{N}$, such that for all $i \geq i_0$, $q_i = \hat{q}$; hence, in view of (3.8), suppose there exists an $\varepsilon > 0$ such that $\theta_i \leq -\varepsilon$ for all $i \geq i_0$. It now follows from the properties of the algorithm defined by (3.3f), (3.3g) (see [13]), that there are two possibilities: either $\psi_{\hat{q}}(x_i) > 0$ for all $i \geq i_0$, or there exists an $i_1 \geq i_0$ such that $\psi_{\hat{q}}(x_i) \leq 0$ for all $i \geq i_1$. In the former case, $\psi_{\hat{q}}(x_{i+1}) - \psi_{\hat{q}}(x_i) \leq \alpha \lambda_i \theta_i$ for all $i \geq i_0$. In the latter case, $\psi_{\hat{q}}^0(x_{i+1}) - \psi_{\hat{q}}^0(x_i) \leq \alpha \lambda_i \theta_i$. Making use of (3.4a) and the assumption that $\psi^0(\cdot)$ is bounded from below, we conclude that $\psi_{\hat{q}}^0(\cdot)$ is also bounded from below. Hence we obtain that either

$$-\infty < \sum_{i=i_0}^{\infty} [\psi_{\hat{q}}(x_{i+1}) - \psi_{\hat{q}}(x_i)] \leq \sum_{i=i_0}^{\infty} \alpha \lambda_i \theta_i,$$

or that

$$-\infty < \sum_{i=i_0}^{\infty} [\psi_{\hat{q}}^0(x_{i+1}) - \psi_{\hat{q}}^0(x_i)] \leq \sum_{i=i_0}^{\infty} \alpha \lambda_i \theta_i.$$

In either event, we are led to the conclusion that $\sum_{i=i_0}^{\infty} \alpha \lambda_i \theta_i > -\infty$. Since $\|h_i\|^2 \leq 2(-\theta_i)$, we conclude, applying the reasoning used in proving Lemma 2.2, that $\{x_i\}_{i=0}^{\infty}$ is a Cauchy sequence in \mathbb{R}^n . Since by Lemma 3.2 in [13] (for algorithm (3.3f), (3.3g)), any accumulation point x^* of $\{x_i\}_{i=0}^{\infty}$ satisfies $\theta_{\hat{q}}(x^*) = 0$, it follows that $\theta_{\hat{q}}(x_i) \rightarrow 0$, which contradicts our assumption. Therefore $\lim_{i \rightarrow \infty} \theta_i = 0$. It now follows from (3.8) and the fact that $q_i \geq q_{i-1}$ that $q_i \rightarrow \infty$ as $i \rightarrow \infty$. \square

THEOREM 3.1. *Suppose that Assumption 3.1 holds, that $\psi^0(\cdot)$ is bounded from below, and that there exists a constant $M \in [1, \infty)$ such that for all $x \in \mathbb{R}^n$, $z \in \mathbb{R}^n$, $j \in \mathbf{L}$, and $q \in \mathbb{N}$,*

$$(3.10) \quad \left\langle z, \frac{\partial^2 \phi_q^j(x, y_j)}{\partial x^2} z \right\rangle \leq M \|z\|^2 \quad \text{for } y_j \in \mathbf{Y}_j.$$

Then any accumulation point \hat{x} of the sequence of iterates $\{x_i\}_{i=0}^{\infty}$ generated by Algorithm 3.1 satisfies $\theta(\hat{x}) = 0$.

Proof. It follows from the definitions of $F_{z,q}(\cdot)$ and $F_z(\cdot)$, and Assumption 3.1(i) that

$$(3.11) \quad |F_{z,q}(x) - F_z(x)| \leq (2 + \gamma)K/q^\tau \quad \text{for all } x \in \mathbb{R}^n, \quad z \in \mathbb{R}^n, \quad q \in \mathbb{N}.$$

Next, we observe the i th iteration of Algorithm 3.1 consists of one iteration of the PPP minimax algorithm on the problem $\min_{x \in \mathbb{R}^n} F_{x_i, q_i}(x)$ starting with x_i . Hence, replacing $\psi_{q_i}(\cdot)$, $\psi(\cdot)$ by $F_{x_i, q_i}(\cdot)$, $F_{x_i}(\cdot)$, we conclude from the proof of Theorem 2.1 that there exists an i_0 such that for all $i \geq i_0$,

$$(3.12) \quad F_{x_i}(x_{i+1}) - F_{x_i}(x_i) \leq \alpha \beta \theta_i / M + 2(2 + \gamma)K/q_i^\tau \leq -\frac{\alpha \beta D^{1/\sigma}}{2M(q_i^\tau)^{1/\sigma}}.$$

It now follows from the definition of $F_{x_i}(\cdot)$, (3.3c), that for all $i \geq i_0$,

$$(3.13a) \quad \psi^0(x_{i+1}) - \psi^0(x_i) - \gamma \psi_+(x_i) \leq -\alpha \beta \theta_i / M + (4 + 2\gamma)K/q_i^\tau \leq -\frac{\alpha \beta D^{1/\sigma}}{2M(q_i^\tau)^{1/\sigma}} < 0,$$

$$(3.13b) \quad \psi(x_{i+1}) - \psi_+(x_i) \leq \alpha \beta \theta_i / M + (4 + 2\gamma)K/q_i^\tau \leq -\frac{\alpha \beta D^{1/\sigma}}{2M(q_i^\tau)^{1/\sigma}} < 0.$$

We must consider two cases.

Case (i). There exists an integer $i_1 > i_0$ such that $\psi(x_{i_1}) \leq 0$. It then follows from (3.13b) that $\psi(x_i) \leq 0$ for all $i \geq i_1$ and from (3.13a) that

$$(3.14) \quad \psi^0(x_{i+1}) - \psi^0(x_i) \leq -\alpha \beta \theta_i / M + (4 + 2\gamma)K/q_i^\tau \leq -\frac{\alpha \beta D^{1/\sigma}}{2M(q_i^\tau)^{1/\sigma}} \quad \text{for all } i \geq i_1.$$

Hence, by the same reasoning used in the proof of Theorem 2.1, we conclude that $\sum_{k=i_1}^{\infty} \theta_k > -\infty$, which leads to the desired result.

Case (ii). $\psi(x_i) > 0$ for all $i > i_0$. It then follows from (3.13b) that for all $i > i_0$,

$$(3.15) \quad \psi(x_{i+1}) - \psi(x_i) \leq -\alpha \beta \theta_i / M + (4 + 2\gamma)K/q_i^\tau \leq -\frac{\alpha \beta D^{1/\sigma}}{2M(q_i^\tau)^{1/\sigma}}.$$

Since $\psi(x_i) > 0$ of all $i \geq i_0$, the reasoning used in the proof of Theorem 2.1 leads to the conclusion that $\sum_{k=i_0}^{\infty} \theta_k > -\infty$. Hence the desired result follows from Lemma 3.1(iii). \square

To establish the linear convergence of Algorithm 3.1, we need the following assumption and results, which we borrow from [13].

Assumption 3.2. (i) There exist $0 < m \leq 1 \leq M < \infty$ such that for all $x \in \mathbb{R}^n$, $z \in \mathbb{R}^n$, and $j \in \mathbf{L}$,

$$(3.16a) \quad m \|z\|^2 \leq \left\langle z, \frac{\partial^2 \phi_q^j(x, y_j)}{\partial x^2} z \right\rangle \leq M \|z\|^2 \quad \text{for all } y_j \in \mathbf{Y}_j, \quad q \in \mathbb{N},$$

$$(3.16b) \quad m \|z\|^2 \leq \left\langle z, \frac{\partial^2 \phi^j(x, y_j)}{\partial x^2} z \right\rangle \leq M \|z\|^2 \quad \text{for all } y_j \in \mathbf{Y}_j.$$

(ii) The set $\{x | \psi(x) < 0\}$ is not empty.

LEMMA 3.3 (Lemma 4.2, Lemma 4.3 in [13]). *Suppose that Assumption 3.2 holds.*

Then

- (i) *Problem (3.1a) has a unique solution \hat{x} ,*
- (ii) *\hat{x} is the unique zero of $\theta(\cdot)$.*
- (iii) *Let $\underline{\mu}^0 \triangleq \min \{\mu^0 | \mu \in L(\hat{x})\}$ and $\bar{\mu}^0 \triangleq \max \{\mu^0 | \mu \in L(\hat{x})\}$, where*

$$(3.17a) \quad L(\hat{x}) \triangleq \left\{ \mu = (\mu^0, \mu^1, \dots, \mu^l) \mid 0 \in \sum_{j=0}^l \mu^j \partial \psi^j(\hat{x}), \right. \\ \left. \sum_{j=1}^l \mu^j \psi^j(\hat{x}) = 0, \sum_{j=0}^l \mu^j = 1, \mu^j \geq 0 \right\}.$$

Then $0 < \underline{\mu}^0 \leq \bar{\mu}^0 \leq 1$.

(iv) *For all $x \in \mathbb{R}^n$,*

$$(3.17b) \quad \bar{\mu}^0 [\psi^0(\hat{x}) - \psi^0(x)] \leq (1 - \bar{\mu}^0) \psi_+(x).$$

THEOREM 3.2. *Suppose that Assumptions 3.1 and 3.2 hold. Then any sequence of iterates $\{x_i\}_{i=0}^\infty$ generated by Algorithm 3.1 converges to \hat{x} , and there exists an interger i_0 such that either*

(i) *$\psi(x_i) \leq 0$ for all $i \geq i_0$ and*

$$(3.18a) \quad \lim_{i \rightarrow \infty} \frac{\psi^0(x_{i+1}) - \psi^0(\hat{x})}{\psi^0(x_i) - \psi^0(\hat{x})} \leq 1 - \underline{\mu}^0 \alpha \beta m / M, \quad \text{or}$$

(ii) *$\psi(x_i) > 0$ for all $i \geq i_0$ and*

$$(3.18b) \quad \lim_{i \rightarrow \infty} \frac{\psi(x_{i+1})}{\psi(x_i)} \leq 1 - \gamma \underline{\mu}^0 \alpha \beta m / M,$$

where $\underline{\mu}^0$ is defined in Lemma 3.3 (iii).

Proof. Let $\hat{x}_{x,q}$, \hat{x}_x be the unique solutions of $\min_{x' \in \mathbb{R}^n} F_{x,q}(x')$ and $\min_{x' \in \mathbb{R}^n} F_x(x')$, respectively. First, we show that both $\{x_i\}_{i=0}^\infty$ and $\{\hat{x}_{x_i}\}_{i=0}^\infty$ converge to \hat{x} . It follows from (3.13b) that $\psi_+(x_{i+1}) \leq \psi_+(x_i)$ when i is sufficiently large. Since $\psi(\cdot)$ has bounded level sets, the sequence $\{x_i\}_{i=0}^\infty$ constructed by Algorithm 3.1 is bounded. It therefore follows from Lemma 3.3(i)-(ii) and Theorem 3.1 that $x_i \rightarrow \hat{x}$ as $i \rightarrow \infty$. Hence, making use of the fact that $F_{x_i}(\hat{x}_{x_i}) \leq F_{x_i}(x_i) = 0$, we deduce that $\{\hat{x}_{x_i}\}_{i=0}^\infty$ is bounded, and that because $F_x(x')$ is continuous in (x, x') any accumulation point \tilde{x} of $\{\hat{x}_{x_i}\}_{i=0}^\infty$, must satisfy $F_{\tilde{x}}(\tilde{x}) \leq 0$, which implies that $\tilde{x} = \hat{x}$. Therefore, $\hat{x}_{x_i} \rightarrow \hat{x}$ as $i \rightarrow \infty$.

Now the i th iteration of Algorithm 3.1 consists of one iteration of the PPP minimax algorithm on $\min_{x \in \mathbb{R}^n} F_{x_i, q_i}(x)$, starting with x_i . Therefore, replacing $\psi_{q_i}(\cdot)$, $\psi(\cdot)$, \hat{x}_{q_i} , and \hat{x} by $F_{x_i, q_i}(\cdot)$, $F_{x_i}(\cdot)$, \hat{x}_{x_i, q_i} , and \hat{x}_{x_i} , respectively, in the proof of Theorem 2.1, we conclude that there exists an integer i_1 such that for all $i \geq i_1$,

$$(3.19a) \quad F_{x_i}(x_{i+1}) - F_{x_i}(\hat{x}_{x_i}) \leq [1 - \nu_i][F_{x_i}(x_i) - F_{x_i}(\hat{x}_{x_i})],$$

where

$$(3.19b) \quad \nu_i \triangleq \frac{\alpha\beta m}{M} - \frac{8(2+\gamma)KM}{D^{1/\sigma}(q_i^\tau)^{(\sigma-1)/\sigma}} > 0.$$

Since $F_{x_i}(x_i) = 0$, we get

$$(3.20) \quad F_{x_i}(x_{i+1}) \leq \nu_i F_{x_i}(\hat{x}_{x_i}).$$

Now, since \hat{x}_i is the minimizer of $F_{x_i}(\cdot)$, there exists a multiplier $\hat{\mu}_i = (\hat{\mu}_i^0, \hat{\mu}_i^1, \dots, \hat{\mu}_i^l)$ such that

$$(3.21a) \quad 0 \in \sum_{j=0}^l \hat{\mu}_i^j \partial \psi^j(\hat{x}_{x_i}),$$

$$(3.21b) \quad \hat{\mu}_i^0 [\psi^0(\hat{x}_{x_i}) - \psi^0(x_i) - \gamma \psi_+(x_i)] + \sum_{j=1}^l \hat{\mu}_i^j [\psi^j(\hat{x}_{x_i}) - \psi_+(x_i)] = F_{x_i}(\hat{x}_{x_i}),$$

$$(3.21c) \quad \sum_{j=0}^l \hat{\mu}_i^j = 1, \quad \hat{\mu}_i^j \geq 0 \quad \text{for } j \in L.$$

Since both sequences $\{x_i\}_{i=0}^\infty$ and $\{\hat{x}_{x_i}\}_{i=0}^\infty$ converge to \hat{x} , any accumulation point of $\{\hat{\mu}_i\}_{i=0}^\infty$ is in $L(\hat{x})$. Hence

$$(3.22) \quad \underline{\mu}^0 \leq \liminf_{i \rightarrow \infty} \hat{\mu}_i^0 \leq \overline{\lim}_{i \rightarrow \infty} \hat{\mu}_i^0 \leq \bar{\mu}^0.$$

Let $\tilde{F}_i(\cdot)$ be defined by

$$(3.33) \quad \tilde{F}_i(x) = \hat{\mu}_i^0 [(\psi^0(x) - \psi^0(x_i) - \gamma \psi_+(x_i))] + \sum_{j=1}^l \hat{\mu}_i^j [\psi^j(x) - \psi_+(x_i)].$$

It follows from (3.21a)–(3.21c) that $0 \in \partial \tilde{F}_i(\hat{x}_{x_i})$. Since $\tilde{F}_i(\cdot)$ is strictly convex, \hat{x}_{x_i} must be its unique minimizer. Hence, making use of (3.21c) and the fact that $\tilde{F}_i(\hat{x}_{x_i}) = F_{x_i}(\hat{x}_{x_i})$ and that $\psi(\hat{x}) \leq 0$, we obtain that

$$(3.24) \quad F_{x_i}(\hat{x}_{x_i}) \leq \tilde{F}_i(\hat{x}) \leq \hat{\mu}_i^0 [(\psi^0(\hat{x}) - \psi^0(x_i))] - [1 + (\gamma - 1)\hat{\mu}_i^0] \psi_+(x_i).$$

Combining (3.20) and (3.24) and rearranging terms, we get

$$(3.25a) \quad \psi^0(x_{i+1}) - \psi^0(\hat{x}) \leq (1 - \hat{\mu}_i^0 \nu_i) [\psi^0(x_i) - \psi^0(\hat{x})] + [\gamma - (1 + (\gamma - 1)\hat{\mu}_i^0) \nu_i] \psi_+(x_i),$$

$$(3.25b) \quad \psi(x_{i+1}) \leq \hat{\mu}_i^0 \nu_i [\psi^0(\hat{x}) - \psi^0(x_i)] + (1 - (1 + (\gamma - 1)\hat{\mu}_i^0) \nu_i) \psi_+(x_i).$$

By Lemma 3.3, $q_i \rightarrow \infty$ as $i \rightarrow \infty$. Hence we deduce from (3.19b) that

$$(3.26) \quad \lim_{i \rightarrow \infty} \nu_i = \frac{\alpha\beta m}{M}.$$

Now, it was shown in the proof of Theorem 3.1 that there exists an integer $i_0 > i_1$ such that either (i) $\psi(x_i) \leq 0$ for all $i \geq i_0$, or (ii) $\psi(x_i) > 0$ for all $i \geq i_0$. In the former case, (3.18a) follows from (3.22), (3.25a), (3.26), and the fact that $\psi_+(x_i) = 0$ for $i \geq i_0$. In the latter case, we obtain from (3.25b) and (3.17b) that

$$(3.27) \quad \begin{aligned} \psi(x_{i+1}) &\leq [\hat{\mu}_i^0 \nu_i (1 - \bar{\mu}^0) / \bar{\mu}^0 + 1 - (1 + (\gamma - 1)\hat{\mu}_i^0) \nu_i] \psi_+(x_i) \\ &= [1 + (\hat{\mu}_i^0 / \bar{\mu}^0 - 1 - \gamma \hat{\mu}_i^0) \nu_i] \psi(x_i). \end{aligned}$$

Hence (3.18b) follows from (3.22), (3.26), and (3.27). \square

For comparison, we reproduce the rate-of-convergence theorem for the USFD algorithm described in [13].

THEOREM 3.3. *Suppose that the relevant part of Assumption 3.2 holds. Then any sequence of iterates $\{x_i\}_{i=0}^\infty$ generated by the USFD algorithm, defined by (3.3e)–(3.3g), converges to \hat{x} , and there exists an integer i_0 such that either (i) $\psi(x_i) \leq 0$ for all $i \geq i_0$ and*

$$(3.28a) \quad \lim_{i \rightarrow \infty} \frac{\psi^0(x_{i+1}) - \psi^0(\hat{x})}{\psi^0(x_i) - \psi^0(\hat{x})} \leq 1 - \underline{\mu}^0 \alpha \beta m / M,$$

or (ii) $\psi(x_i) > 0$ for all $i \geq i_0$ and

$$(3.28b) \quad \lim_{i \rightarrow \infty} \frac{\psi(x_{i+1})}{\psi(x_i)} \leq 1 - \gamma \underline{\mu}^0 \alpha \beta m / M,$$

where $\underline{\mu}^0$ is defined in Lemma 3.3(iii).

Thus we see that Algorithm 3.1 has the same rate of convergence as the USFD algorithm, and hence, by the same arguments used at the end of § 2, we conclude that using adaptive discretization in the form of Algorithm 3.1 should result in savings in computing time over the use of the USFD algorithm on a single high-precision approximation to the original problem.

4. Optimal control problems. Finally, we turn to unconstrained optimal control problems. The most natural space for establishing the differentiability of solutions of differential equations with respect to controls is L_∞ . However, this is not a natural space for extending finite-dimensional algorithms, defined on the Hilbert space \mathbb{R}^n , to optimal control problems. The natural space for this purpose is L_2 , but in L_2 , solutions of differential equations are Fréchet differentiable with respect to controls in L_2 only under a *growth condition* [2], [20]. Fortunately, since in the solution of optimal control problems, control sequences remain bounded in the L_∞ -norm, imposition of a growth condition, as we will shortly describe, amounts to only a technical artifice.

We will consider unconstrained optimal control problems of the form

$$(4.1a) \quad \text{OCP: } \min_{u \in G} c(u) = g(\bar{x}(u)),$$

where $g: \mathbb{R}^n \rightarrow \mathbb{R}$, $G \triangleq L_2^m[0, T]$, T is a given time period, and $\bar{x}: G \rightarrow \mathbb{R}^n$ is defined by $\bar{x}(u) \triangleq x(T, u, x_0)$, with $x(\cdot, u, x_0)$ the solution of the differential equation

$$(4.1b) \quad \dot{x}(t) = f(x(t), u(t), t), \quad t \in [0, T], \quad x(0) = x_0,$$

where x_0 is a given vector in \mathbb{R}^n .

In Assumption 4.1 below, we postulate local Lipschitz continuity conditions that are independent of bounds on the control. We justify this assumption as follows. We may assume that there exists a compact set $U \subset \mathbb{R}^m$, such that the control sequences $\{u_i(\cdot)\}_{i=0}^\infty$, constructed by our algorithm, take values in U . Hence there exists a bound $b < \infty$ such that for all $u \in U$, $|u^i| \leq b$, $i = 1, 2, \dots, m$. Consequently, we may assume without loss of generality that the function f has the form $f(x, u, t) = \tilde{f}(x, \text{SAT}(u), t)$, where $\text{SAT}: \mathbb{R}^m \rightarrow \mathbb{R}^m$ is such that

$$\text{SAT}(u) = (\text{sat}(u^1), \text{sat}(u^2), \dots, \text{sat}(u^m))$$

and for all $z \in \mathbb{R}$, $\text{sat}(z) = z$ if $|z| \leq 2b$ and $\text{sat}(z) = 2b \operatorname{sgn}(z)(1 - e^{-(|z|-2b)})$ otherwise, and $\tilde{f}(\cdot, \cdot, \cdot)$ satisfies standard assumptions. It is easy to see that Assumption 4.1 is satisfied under such a *growth condition* on the right-hand side of (4.1b).

We will assume that functions $g: \mathbb{R}^n \rightarrow \mathbb{R}$ and $f: \mathbb{R}^n \times \mathbb{R}^m \times [0, T] \rightarrow \mathbb{R}^n$ have the following properties.

Assumption 4.1. (i) The function $g(\cdot)$ and its gradient $\nabla g(\cdot)$ are locally Lipschitz continuous.

(ii) The function $f(\cdot, \cdot, \cdot)$ and its partial derivatives $\partial f(\cdot, \cdot, \cdot)/\partial x$ and $\partial f(\cdot, \cdot, \cdot)/\partial u$ are locally Lipschitz continuous, with the Lipschitz constant independent of the magnitude of the control u .

(iii) There exists a constant K_1 such that for all $x \in \mathbb{R}^n$, $u \in L_2^m[0, T]$, and $t \in [0, T]$,

$$(4.2a) \quad \|f(x, u, t)\| \leq K_1[\|x\| + 1],$$

$$(4.2b) \quad \left\| \frac{\partial f(x, u, t)}{\partial x} \right\| \leq K_1[\|x\| + 1],$$

$$(4.2c) \quad \left\| \frac{\partial f(x, u, t)}{\partial u} \right\| \leq K_1[\|x\| + 1].$$

The following result can be obtained using the implicit function theorem in Banach spaces, as shown in [2], [20].

LEMMA 4.1. *Suppose that Assumption 4.1 is satisfied. Then*

(i) *The differential equation (4.1b) has a unique solution for every $u \in G$.*

(ii) *The functions $\bar{x}(\cdot)$ and $c(\cdot)$ are continuously Fréchet differentiable on G .*

Since the functions $\bar{x}(\cdot)$ and $c(\cdot)$ are continuously Fréchet differentiable on G , there exist continuous functions, a “Jacobian” $(n \times m)$ -matrix-valued function $(\partial \bar{x}/\partial u)(t)$, $t \in [0, t]$, and a “gradient” $\nabla c: G \rightarrow G$, such that

$$(4.3a) \quad \lim_{\substack{\delta u \in G \\ \|\delta u\|_2 \rightarrow 0}} \frac{\left\| \bar{x}(u + \delta u) - \bar{x}(u) - \int_0^T \frac{\partial \bar{x}(u)}{\partial u}(t) \delta u(t) dt \right\|}{\|\delta u\|_2} = 0,$$

$$(4.3b) \quad \lim_{\substack{\delta u \in G \\ \|\delta u\|_2 \rightarrow 0}} \frac{\left| c(u + \delta u) - c(u) - \int_0^T \langle \nabla c(u)(t), \delta u(t) \rangle dt \right|}{\|\delta u\|_2} = 0.$$

The following result can be deduced from basic differential equation theory (see [2], [20]) and Lemma 4.1.

LEMMA 4.2. *Suppose that Assumption 4.1 is satisfied. Then (i) For every $u \in G$ and $t \in [0, T]$,*

$$(4.4a) \quad \frac{\partial \bar{x}(u)}{\partial u}(t) = \Phi_u(T, t) \frac{\partial f(x(u, t), u(t), t)}{\partial u},$$

$$(4.4b) \quad \nabla c(u)(t) = \left[\frac{\partial \bar{x}(u)}{\partial u}(t) \right]^T \nabla g(\bar{x}(u)),$$

where $\Phi_u(t, \tau)$ is the state transition matrix for the linear differential equation

$$(4.4c) \quad \dot{y}(t) = \frac{\partial f(x(t, u), u(t), t)}{\partial x} y(t) \text{ on } t \in [0, T];$$

(ii) *There exists a K_2 such that for all $u, \delta u \in G$,*

$$(4.4d) \quad \|\bar{x}(u)\| \leq K_2, \quad \left\| \frac{\partial \bar{x}(u)}{\partial u} \right\|_\infty \leq K_2, \quad \|\nabla c(u)\|_\infty \leq K_2,$$

$$(4.4e) \quad \|\bar{x}(u + \delta u) - \bar{x}(u)\| \leq K_2 \|\delta u\|_2,$$

$$(4.4f) \quad \left\| \bar{x}(u + \delta u) - \bar{x}(u) - \int_0^T \left[\frac{\partial \bar{x}(u)}{\partial u}(t) \right] \delta u(t) dt \right\| \leq K_2 \|\delta u\|_2^2.$$

As a first step toward the numerical solution of the infinite-dimensional problem **OCP**, we define a sequence of finite-dimensional subspaces G_q of G , $q \in \mathbb{N}$. Thus, for any $q \in \mathbb{N}$, let $\Delta_q \triangleq T/2^q$, and let

$$G_q \triangleq G \cap \{u \mid u(t) = u^j \in \mathbb{R}^m, \quad t \in [j\Delta_q, (j+1)\Delta_q), j = 0, 1, \dots, 2^q - 1\}.$$

Next, for any $q \in \mathbb{N}$, and any $u \in G_q$, let $\bar{x}_q(u)$ be an approximation to $\bar{x}(u)$, obtained by solving the differential equation (4.1b) by means of a numerical method, such as the Euler–Cauchy method, the modified Euler method, the Runge–Kutta method, and others. We can now define a family of finite-dimensional approximating problems, parametrized by the discretization parameter $q \in \mathbb{N}$

$$(4.5) \quad \mathbf{OCP}_q : \min_{u \in G_q} c_q(u),$$

where $c_q(u) \triangleq g(\bar{x}_q(u))$. We will assume that $\bar{x}_q(\cdot)$ approximates $\bar{x}(u)$ in the following sense.

Assumption 4.2. (i) There exists a $\tau \in (0, \infty)$ such that there exist constants K_3 , such that for any $q \in \mathbb{N}$,

$$(4.6a) \quad |x(u) - \bar{x}_q(u)| \leq K_3/(2^q)^\tau \quad \text{for all } u \in G_q,$$

(ii) $\bar{x}_q(\cdot)$ is continuously Fréchet differentiable on G_q . We will denote its “Jacobian” by $(\partial \bar{x}_q(u)/\partial u)(t)$.

(iii) For any $\varepsilon \in (0, \infty)$, there exists a $\hat{q} \in \mathbb{N}$ such that for all $q \geq \hat{q}$ and $u \in G_q$,

$$(4.6b) \quad \left\| \frac{\partial \bar{x}_q(u)}{\partial u} - \frac{\partial \bar{x}(u)}{\partial u} \right\|_\infty \leq \varepsilon;$$

(iv) There exists $K_4 \in (0, \infty)$ such that for all $q \in \mathbb{N}$ and all $u, \delta u \in G_q$,

$$(4.6c) \quad \|\bar{x}_q(u + \delta u) - \bar{x}_q(u)\| \leq K_4 \|\delta u\|_2,$$

$$(4.6d) \quad \left\| \bar{x}_q(u + \delta u) - \bar{x}_q(u) - \int_0^T \left[\frac{\partial \bar{x}_q(u)}{\partial u}(t) \right] \delta u(t) dt \right\| \leq K_4 \|\delta u\|_2^2. \quad \square$$

Remark 4.1. Referring to [7], we see that when the Euler–Cauchy method is used to define $\bar{x}_q(u)$, Assumption 4.2 is satisfied with $\tau = 1$. It is easy to show that when the modified Euler method or Runge–Kutta method are used to define $\bar{x}_q(u)$, Assumption 4.2 is satisfied with $\tau = 3$ and $\tau = 5$, respectively.

LEMMA 4.3. Suppose that Assumptions 4.1 and 4.2 are satisfied. Then (i) The function $c_q(\cdot)$ is continuously Fréchet differentiable on G_q , and for any $u \in G_q$ and $t \in [0, T]$,

$$(4.7a) \quad \nabla c_q(u)(t) = \left[\frac{\partial \bar{x}_q(u)}{\partial u}(t) \right]^T \nabla g(\bar{x}_q(u));$$

(ii) There exists a $K_5 \in (0, \infty)$ such that for $q \in \mathbb{N}$ and all $u \in G_q$,

$$(4.7b) \quad |c(u) - c_q(u)| \leq K_5/(2^q)^\tau.$$

Proof. Part (i) follows from the Assumption 4.2(ii), (4.6c), and the local Lipschitz continuity of $\nabla g(\cdot)$. Inequality (4.7b) follows from Assumption 4.2(i), (4.4d), and the local Lipschitz continuity of $g(\cdot)$. \square

Since G_q is isomorphic to \mathbb{R}^{2^q} , each problem **OCP**_q defined by (4.5) can be solved by the Armijo gradient method [1], which uses two parameters $\alpha, \beta \in (0, 1)$ and which for **OCP**_q constructs iterates according to the rule

$$(4.8a) \quad u_{i+1} = u_i - \lambda_i \nabla c_q(u_i),$$

with

$$(4.8b) \quad \lambda_i = \max \{ \beta^k \mid k \in \mathbb{N}, c_q(u_i - \beta^k \nabla c_q(u_i)) - c_q(u_i) \leq -\alpha \beta^k \|\nabla c_q(u_i)\|_2^2 \}.$$

We can now state an adaptive discretization scheme, based on the Armijo gradient method (4.8a), (4.8b), for solving the problem **OCP**.

Adaptive Discretization Algorithm 4.1 (for OCP):

Data: $u_0 \in G_{q_{-1}}$, $q_{-1} \in \mathbb{N}$, $\alpha \in (0, 1)$, $\beta \in (0, 1)$, $D > 0$, and $\sigma > 1$.

Step 0: Set $i = 0$.

Step 1: Compute $q_i \in \mathbb{N}$ and $\nabla c_{q_i}(u_i)$ such that $q_i \geq q_{i-1}$ and

$$(4.9a) \quad D / (2^{q_i})^\tau \leq [\|\nabla c_{q_i}(u_i)\|_2^2]^\sigma.$$

Step 2: Set $h_i = -\nabla c_{q_i}(u_i)$, $\theta_i = -\|\nabla c_{q_i}(u_i)\|_2^2$ and compute the stepsize λ_i :

$$(4.9b) \quad \lambda_i = \max \{ \beta^k \mid k \in \mathbb{N}, c_{q_i}(u_i + \beta^k h_i) - c_{q_i}(u_i) \leq \alpha \beta^k \theta_i \}.$$

Step 3: Set $u_{i+1} = u_i + \lambda_i h_i$, replace i by $i + 1$, and go to Step 1.

Remark 4.2. It follows from Assumptions 4.2(i) and 4.2(iii), (4.4b), and (4.7a), that whenever $\nabla c(u_i) \neq 0$ (i.e., u_i does not satisfy a first-order optimality condition for the problem **OCP**), Step 1 of Algorithm 4.1 yields a finite q_i . For simplicity, in the rest of this section we will assume that Algorithm 4.1 does not construct a u_i such that $\nabla c(u_i) = 0$, for any finite i .

LEMMA 4.4. Suppose that Assumption 4.1 is satisfied, that $g(\cdot)$ is bounded from below, and that the sequence of controls $\{u_i\}_{i=0}^\infty$ and the corresponding sequence of discretization parameters $\{q_i\}_{i=0}^\infty$ are constructed by Algorithm 4.1. Then $q_i \rightarrow \infty$ as $i \rightarrow \infty$.

Proof. Suppose that $q_i \rightarrow \infty$ as $i \rightarrow \infty$ does not hold. Then, using the same reasoning as in the proof of Lemma 2.2 and the fact that $\|h_i\|_2^2 = -\theta_i$, we conclude that there exist i_0 and $\hat{q} \in \mathbb{N}$ such that $q_i = \hat{q}$ for all $i \geq i_0$, and that $\{u_i\}_{i=0}^\infty$ is a Cauchy sequence in $G_{\hat{q}}$. Since for any accumulation point \hat{u} , of a sequence $\{u_i\}_{i=0}^\infty$ constructed by the Armijo gradient method (4.8a), (4.8b) in the finite-dimensional space, $G_{\hat{q}}$, $\nabla c_{\hat{q}}(\hat{u}) = 0$ (see [1]), $\theta_{\hat{q}}(u_i) = -\|\nabla c_{\hat{q}}(u_i)\|_2^2 \rightarrow 0$, which contradicts the test (4.9a). Thus we must have that $q_i \rightarrow \infty$ as $i \rightarrow \infty$. \square

THEOREM 4.1. Suppose that Assumptions 4.1 and 4.2 are satisfied, that $g(\cdot)$ is twice continuously differentiable and bounded from below, and that there exists a constant $M_g \in (0, \infty)$ such that for all $x \in \mathbb{R}^n$, $z \in \mathbb{R}^n$,

$$(4.10) \quad \left\langle z, \frac{\partial^2 g(x)}{\partial x^2} z \right\rangle \leq M_g \|z\|^2.$$

Then any accumulation point $\hat{u} \in G$ of a sequence of controls $\{u_i\}_{i=0}^\infty$ generated by Algorithm 4.1 satisfies $\nabla c(\hat{u}) = 0$.

Proof. Making use of (4.10), (4.7a), (4.6c), (4.6d), and the fact that $h_i = -\nabla c_{q_i}(u_i)$ and that $\theta_i = -\|h_i\|_2^2$, we obtain that for all $\lambda \in [0, 1]$,

$$\begin{aligned} c_q(u_i + \lambda h_i) - c_q(u_i) &= g(\bar{x}_{q_i}(u_i + \lambda h_i)) - g(\bar{x}_{q_i}(u_i)) \\ &\leq \langle \nabla g(\bar{x}_{q_i}(u_i)), (\bar{x}_{q_i}(u_i + \lambda h_i) - \bar{x}_{q_i}(u_i)) \rangle \\ &\quad + \frac{M_g}{2} \|\bar{x}_{q_i}(u_i + \lambda h_i) - \bar{x}_{q_i}(u_i)\|^2 \\ (4.11) \quad &\leq \left\langle \nabla g(\bar{x}_{q_i}(u_i)), \int_0^T \frac{\partial \bar{x}_{q_i}(u_i)}{\partial u}(t) \lambda h_i(t) dt \right\rangle \\ &\quad + K_4 \|\nabla g(\bar{x}_{q_i}(u_i))\| \|\lambda h_i\|_2^2 + \frac{M_g}{2} (K_4 \|\lambda h_i\|_2)^2 \\ &= \lambda \theta_i - \lambda^2 [K_4 \|\nabla g(\bar{x}_{q_i}(u_i))\| + M_g K_4^2 / 2] \theta_i. \end{aligned}$$

Since $\bar{x}_{q_i}(u_i)$ is bounded and $\nabla g(\cdot)$ is continuous, there exists a $K_6 \in (1, \infty)$ such that

$$(4.12) \quad c_q(u_i + \lambda h_i) - c_q(u_i) \leq (\lambda - \lambda^2 K_6) \theta_i = \alpha \lambda \theta_i + \lambda(1 - \alpha - K_6 \lambda) \theta_i, \\ \forall i \in \mathbb{N} \quad \forall \lambda \in [0, 1].$$

Hence (4.9b) is satisfied with

$$(4.13) \quad \lambda_i \leq (1 - \alpha) \beta / K_6,$$

and thus

$$(4.14) \quad c_{q_i}(u_{i+1}) - c_{q_i}(u_i) \leq \alpha \lambda_i \theta_i \leq \alpha(1 - \alpha) \beta \theta_i / K_6.$$

It therefore follows from (4.7b) that

$$(4.15) \quad c(u_{i+1}) - c(u_i) \leq \alpha(1 - \alpha) \beta \theta_i / K_6 + 2K_5 / (2^{q_i})^\tau.$$

Resorting to the reasoning used in the proof of Theorem 2.1, with $\psi(\cdot)$, x_i , q_i , M , and K replaced by $c(\cdot)$, u_i , 2^{q_i} , $K_6/(1 - \alpha)$, and K_5 , we can show that $\sum_{k=0}^{\infty} \theta_k > -\infty$. Hence $\theta_i \rightarrow 0$ as $i \rightarrow \infty$. Now, for all $i \in \mathbb{N}$,

$$(4.16) \quad \|\nabla c(\hat{u})\|_2 \leq \|\nabla c(\hat{u}) - \nabla c(u_i)\|_2 + \|\nabla c(u_i) - \nabla c_{q_i}(u_i)\|_2 + \|\nabla c_{q_i}(u_i)\|_2 \\ \leq \|\nabla c(\hat{u}) - \nabla c(u_i)\|_2 + T^{1/2} \|\nabla c(u_i) - \nabla c_{q_i}(u_i)\|_\infty + (-\theta_i)^{1/2}.$$

Consequently, the desired result follows from the continuity of $\nabla c(\cdot)$, (4.6b), Assumptions 4.1 and 4.2, Lemma 4.3 (specifically, (4.7a)), and the facts that $\theta_i \rightarrow 0$ and $q_i \rightarrow \infty$ as $i \rightarrow \infty$. \square

LEMMA 4.5. *Suppose that Assumptions 4.1 and 4.2 hold and that there exist $0 < m_g < M_g < \infty$ and $0 < m_c < M_c < \infty$ such that for all $x \in \mathbb{R}^n$, $z \in \mathbb{R}^n$, and $u \in G$,*

$$(4.17a) \quad m_g \|z\|^2 \leq \left\langle z, \frac{\partial^2 g(x)}{\partial x^2} z \right\rangle \leq M_g \|z\|^2,$$

$$(4.17b) \quad m_c \|z\|^2 \leq \left\langle z, \left\{ \int_0^T \left[\frac{\partial \bar{x}(u)}{\partial u}(t) \right] \left[\frac{\partial \bar{x}(u)}{\partial u}(t) \right]^T dt \right\} z \right\rangle \leq M_c \|z\|^2.$$

Then, for any $\varepsilon \in (0, m_c)$, there exists a \hat{q} such that, for any $q \geq \hat{q}$ and $u \in G_q$,

$$(4.18a) \quad (m_c - \varepsilon) \|z\|^2 \leq \left\langle z, \left\{ \int_0^T \left[\frac{\partial \bar{x}_q(u)}{\partial u}(t) \right] \left[\frac{\partial \bar{x}_q(u)}{\partial u}(t) \right]^T dt \right\} z \right\rangle \leq (M_c + \varepsilon) \|z\|^2,$$

$$(4.18b) \quad (m_c - \varepsilon) \|\nabla g(\bar{x}_q(u))\|^2 \leq \|\nabla c_q(u)\|_2^2 \leq (M_c + \varepsilon) \|\nabla g(\bar{x}_q(u))\|^2,$$

$$(4.18c) \quad \frac{1}{2M_g(M_c + \varepsilon)} \|\nabla c_q(u)\|_2^2 \leq c_q(u) - g(\hat{x}) \leq \frac{1}{2m_g(m_c - \varepsilon)} \|\nabla c_q(u)\|_2^2,$$

where \hat{x} is the unique minimizer of $g(\cdot)$.

Proof. Inequality (4.18a) follows directly from (4.4d), (4.17b), and Assumption 4.2 (iii), while (4.18b) follows from (4.7a) and (4.18a). Making use of (4.18b) and the fact that for all $x \in \mathbb{R}^n$,

$$(4.19) \quad \frac{1}{2M_g} \|\nabla g(x)\|^2 \leq g(x) - g(\hat{x}) \leq \frac{1}{2m_g} \|\nabla g(x)\|^2,$$

we get (4.18c). \square

Remark 4.3. The matrix

$$\int_0^T \left[\frac{\partial \bar{x}(u)}{\partial u}(t) \right] \left[\frac{\partial \bar{x}(u)}{\partial u}(t) \right]^T dt$$

is the controllability Grammian of the linearization of system (4.1b). When the dynamical system (4.1b) is linear, inequality (4.17b) holds if and only if (4.1b) is completely controllable on $[0, T]$. When the dynamical system (4.1b) is nonlinear, condition (4.17b) is a sufficient condition for the complete controllability of system (4.1b) on $[0, T]$ (see [16]).

THEOREM 4.2. *Suppose that Assumptions 4.1 and 4.2 are satisfied and that there exist $0 < m_g < M_g < \infty$ and $0 < m_c < M_c < \infty$ such that $M_g M_c \geq 2$ and (4.17a), (4.17b) hold. Let \hat{x} be the unique minimizer of $g(\cdot)$. If $\{u_i\}_{i=0}^\infty$ is a sequence of controls generated by Algorithm 4.1, then*

$$(4.20a) \quad (i) \quad \lim_{i \rightarrow \infty} \nabla g(\bar{x}_{q_i}(u_i)) = 0,$$

$$(4.20b) \quad (ii) \quad \lim_{i \rightarrow \infty} c(u_i) = g(\hat{x}),$$

$$(4.20c) \quad (iii) \quad \overline{\lim}_{i \rightarrow \infty} \frac{c(u_{i+1}) - g(\hat{x})}{c(u_i) - g(\hat{x})} \leq 1 - \frac{4(1-\alpha)\alpha\beta m_g m_c}{M_g M_c},$$

$$(iv) \quad \text{The sequence } \{\bar{x}_{q_i}(u_i)\}_{i=0}^\infty \text{ converges } R\text{-linearly to } \hat{x},$$

$$(v) \quad \text{There exists a } \hat{u} \in G \text{ such that } \nabla c(\hat{u}) = 0 \text{ and the sequence } \{\|u_i - \hat{u}\|_\infty\}_{i=0}^\infty \text{ converges } R\text{-linearly to } 0.$$

Proof. Making use of Lemmas 4.4 and 4.5, we conclude that for every $\varepsilon \in (0, m_c)$, there exists an i_ε such that for all $i \geq i_\varepsilon$, (4.18a)–(4.18c) hold for $q = q_i$ and all $u \in G_{q_i}$.

(i) By Theorem 4.1, $\|\nabla c_{q_i}(u_i)\|_2 \rightarrow 0$ as $i \rightarrow \infty$. Hence it follows from (4.18b) that $\nabla g(\bar{x}_{q_i}(u_i)) \rightarrow 0$ as $i \rightarrow \infty$.

(ii) Equation (4.20) follows from (4.7b), (4.18c), and the fact that $\|\nabla c_{q_i}(u_i)\|_2 \rightarrow 0$ and $q_i \rightarrow \infty$ as $i \rightarrow \infty$.

(iii) First, we will obtain a bound on the stepsize λ_i . Making use of (4.6d), (4.18a), and the fact that

$$h_i(t) = - \left[\frac{\partial \bar{x}_{q_i}(u_i)}{\partial u}(t) \right]^T \nabla g(\bar{x}_{q_i}(u_i))$$

and that $\langle v, A^2 v \rangle \leq \|A\| \langle v, Av \rangle$ for all symmetric, positive definite matrices A and vectors v , we obtain that for all $i \geq i_\varepsilon$ and $\lambda \in [0, 1]$,

$$\begin{aligned} \|\bar{x}_{q_i}(u_i + \lambda h_i) - \bar{x}_{q_i}(u_i)\| &\leq \left\| \int_0^T \left[\frac{\partial \bar{x}_{q_i}(u_i)}{\partial u}(t) \right] \lambda h_i(t) dt \right\| + K_4 \|\lambda h_i\|_2^2 \\ &= \lambda \left[\left\langle \nabla g(\bar{x}_{q_i}(u_i)), \left(\int_0^T \left[\frac{\partial \bar{x}_{q_i}(u_i)}{\partial u}(t) \right] \right. \right. \right. \\ &\quad \cdot \left. \left. \left[\frac{\partial \bar{x}_{q_i}(u_i)}{\partial u}(t) \right]^T dt \right)^T \nabla g(\bar{x}_{q_i}(u_i)) \right\rangle \right]^{1/2} + \lambda^2 K_4 \|h_i\|_2^2 \\ (4.21) \quad &\leq \lambda \left[(M_c + \varepsilon) \left\langle \nabla g(\bar{x}_{q_i}(u_i)), \left(\int_0^T \left[\frac{\partial \bar{x}_{q_i}(u_i)}{\partial u}(t) \right] \right. \right. \right. \\ &\quad \cdot \left. \left. \left[\frac{\partial \bar{x}_{q_i}(u_i)}{\partial u}(t) \right]^T dt \right)^T \nabla g(\bar{x}_{q_i}(u_i)) \right\rangle \right]^{1/2} + \lambda^2 K_4 \|h_i\|_2^2 \\ &\leq [(M_c + \varepsilon)^{1/2} + \lambda K_4 \|h_i\|_2] \lambda \|h_i\|_2. \end{aligned}$$

Next, we deduce from (4.17a), (4.6d), (4.7a), (4.21), and the fact that $h_i = -\nabla c_{q_i}(u_i)$, that for all $\lambda \in [0, 1]$,

$$\begin{aligned}
 c_q(u_i + \lambda h_i) - c_q(u_i) &= g(\bar{x}_{q_i}(u_i + \lambda h_i)) - g(\bar{x}_{q_i}(u_i)) \\
 &\leq \langle \nabla g(\bar{x}_{q_i}(u_i)), (\bar{x}_{q_i}(u_i + \lambda h_i) - \bar{x}_{q_i}(u_i)) \rangle \\
 &\quad + \frac{M_g}{2} \|\bar{x}_{q_i}(u_i + \lambda h_i) - \bar{x}_{q_i}(u_i)\|^2 \\
 &\leq -\lambda \|h_i\|_2^2 + \lambda^2 K_4 \|\nabla g(\bar{x}_{q_i}(u_i))\| \|h_i\|_2^2 \\
 &\quad + \frac{M_g}{2} \lambda^2 ((M_c + \varepsilon)^{1/2} + \lambda K_4 \|h_i\|_2)^2 \|h_i\|_2^2.
 \end{aligned}
 \tag{4.22}$$

Since $\|h_i\|_2 \rightarrow 0$ and $\|\nabla g(\bar{x}_{q_i}(u_i))\| \rightarrow 0$ as $i \rightarrow \infty$, it follows from (4.22) that there exists $i'_\varepsilon \geq i_\varepsilon$ such that for $i \geq i'_\varepsilon$ and $\lambda \in [0, 1]$,

$$\begin{aligned}
 c_{q_i}(u_i + \lambda h_i) - c_{q_i}(u_i) &\leq -\lambda \|h_i\|_2^2 + M_g(M_c + 2\varepsilon)\lambda^2 \|h_i\|_2^2/2 \\
 &= \lambda \theta_i - M_g(M_c + 2\varepsilon)\lambda^2 \theta_i/2.
 \end{aligned}
 \tag{4.23}$$

Hence (4.9b) is satisfied with $\lambda_i \geq 2(1 - \alpha)\beta/[M_g(M_c + 2\varepsilon)]$ for all $i \geq i'_\varepsilon$, and thus

$$c_{q_i}(u_{i+1}) - c_{q_i}(u_i) \leq \alpha \lambda_i \theta_i \leq 2\alpha(1 - \alpha)\beta \theta_i/[M_g(M_c + 2\varepsilon)], \quad \forall i \geq i'_\varepsilon.
 \tag{4.24}$$

Combining (4.24) and (4.18c) and rearranging terms, we obtain that for all $i \geq i'_\varepsilon$

$$c_{q_i}(u_{i+1}) - g(\hat{x}) \leq \left[1 - \frac{4(1 - \alpha)\alpha\beta m_g(m_c - \varepsilon)}{M_g(M_c + 2\varepsilon)} \right] [c_{q_i}(u_i) - g(\hat{x})].
 \tag{4.25}$$

Hence it follows from (4.7b) that for $i \geq i'_\varepsilon$,

$$c(u_{i+1}) - g(\hat{x}) \leq \left[1 - \frac{4(1 - \alpha)\alpha\beta m_g(m_c - \varepsilon)}{M_g(M_c + 2\varepsilon)} \right] [c(u_i) - g(\hat{x})] + 2K_5/(2^{q_i})^\tau.
 \tag{4.26}$$

Finally, making use of (4.7b), (4.18c), and the fact that $-\theta_i \geq D^{1/\sigma}/((2^{q_i})^\tau)^{1/\sigma}$, we obtain that for $i \geq i'_\varepsilon$,

$$\begin{aligned}
 2M_g(M_c + \varepsilon)[c(u_i) - g(\hat{x})] &\geq 2M_g(M_c + \varepsilon)[c_{q_i}(u_i) - g(\hat{x})] - 2M_g(M_c + \varepsilon)K_5/(2^{q_i})^\tau \\
 &\geq -\theta_i - 2M_g(M_c + \varepsilon)K_5/(2^{q_i})^\tau \\
 &\geq D^{1/\sigma}/((2^{q_i})^\tau)^{1/\sigma} - 2M_g(M_c + \varepsilon)K_5/(2^{q_i})^\tau.
 \end{aligned}
 \tag{4.27}$$

Since $\sigma > 1$ and by Lemma 4.3 $q_i \rightarrow \infty$ as $i \rightarrow \infty$, we claim that there exists $i''_\varepsilon \geq i'_\varepsilon$ such that for $i \geq i''_\varepsilon$

$$4M_g M_c [c(u_i) - g(\hat{x})] \geq \frac{1}{2} D^{1/\sigma}/((2^{q_i})^\tau)^{1/\sigma}.
 \tag{4.28}$$

Thus (4.20c) follows from (4.26), (4.28), and the arbitrary choice of ε in $(0, m_c)$.

(iv) Note that

$$\| \bar{x}_{q_i}(u_i) - \hat{x} \|^2 \leq \frac{2}{m_g} [g(\bar{x}_{q_i}(u_i)) - g(\hat{x})] = \frac{2}{m_g} [c_{q_i}(u_i) - g(\hat{x})].
 \tag{4.29}$$

This, together with (4.25), leads to the desired result.

(v) Since $\|u_i\|_\infty$ is bounded and, by Lemma 4.4, $q_i \rightarrow \infty$ as $i \rightarrow \infty$, it follows from Assumptions 4.1(iii) and 4.2(ii) and Lemma 4.2(ii) that there exists an $\omega > 0$ such that

for all $i \geq 0$, $\|(\partial \bar{x}_{q_i}(u_i))/\partial u\|_\infty \leq \omega$. Making use of this fact, of (4.19), (4.25), and of the fact that $\lambda_i \leq 1$, we obtain that for $i \geq i'_\varepsilon$,

$$\begin{aligned}
 \|u_{i+1} - u_i\|_\infty &\leq \lambda_i \|h_i\|_\infty \\
 &\leq \lambda_i \left\| \frac{\partial \bar{x}_{q_i}(u_i)}{\partial u} \right\|_\infty \|\nabla g(\bar{x}_{q_i}(u_i))\| \\
 (4.30) \quad &\leq \omega [2M_g[g(\bar{x}_{q_i}(u_i)) - g(\hat{x})]]^{1/2} \\
 &\leq \omega \left[2M_g \left[c_{q_{i'_\varepsilon}}(u_{i'_\varepsilon}) - g(\hat{x}) \right] \left[1 - \frac{4(1-\alpha)\alpha\beta m_g(m_c - \varepsilon)}{M_q(M_c + 2\varepsilon)} \right] \right]^{(1-i'_\varepsilon)/2}.
 \end{aligned}$$

Let

$$\eta \triangleq \left[1 - \frac{4(1-\alpha)\alpha\beta m_g(m_c - \varepsilon)}{M_q(M_c + 2\varepsilon)} \right]^{1/2}.$$

Then, for all $j > 1 \geq i'_\varepsilon$,

$$\begin{aligned}
 \|u_j - u_i\|_\infty &\leq \sum_{k=i}^{j-1} \|u_{k+1} - u_k\|_\infty \\
 (4.31) \quad &\leq \omega [2M_g[c_{q_{i'_\varepsilon}}(u_{i'_\varepsilon}) - g(\hat{x})]]^{1/2} \sum_{k=i}^{\infty} \eta^{k-i'_\varepsilon} \\
 &\leq \omega [2M_g[c_{q_{i'_\varepsilon}}(u_{i'_\varepsilon}) - g(\hat{x})]]^{1/2} \eta^{i-i'_\varepsilon} / (1-\eta).
 \end{aligned}$$

Therefore $\{u_i\}_{i=0}^\infty$ is a Cauchy sequence in $L_\infty^m[0, T]$. Since $L_\infty^m[0, T]$ is a complete space in the L_∞ -norm, there exists a $\hat{u} \in L_\infty^m[0, T]$ such that $u_i \rightarrow \hat{u}$ as $i \rightarrow \infty$ in L_∞ -norm. Let j go to ∞ in (4.31), then for all $i \geq i'_\varepsilon$,

$$(4.32) \quad \|\hat{u} - u_i\|_\infty \leq \omega [2M_g[c_{q_{i'_\varepsilon}}(u_{i'_\varepsilon}) - g(\hat{x})]]^{1/2} \eta^{i-i'_\varepsilon} / (1-\eta).$$

Thus, $\{\|\hat{u} - u_i\|_\infty\}_{i=0}^\infty$ converges to zero R -linearly. Finally, by Theorem 4.1, $\nabla c(\hat{u}) = 0$. \square

For comparison, we present a rate-of-convergence result for the Armijo algorithm (4.8a), (4.8b), as applied to composite functions. The proof of this theorem follows by generalization of the Armijo algorithm rate-of-convergence theorem for affine-composite functions presented in [14].

THEOREM 4.3. *Consider the problem*

$$(4.33a) \quad \min_{u \in \mathbb{R}^N} c(u),$$

where $c(u) \triangleq g(\bar{x}(u))$, with $g: \mathbb{R}^n \rightarrow \mathbb{R}$ twice continuously differentiable function satisfying (4.17a), with $0 < m_g \leq M_g < \infty$, and $\bar{x}: \mathbb{R}^N \rightarrow \mathbb{R}^n$ is a Lipschitz continuously differentiable function such that for some $0 < m_c \leq M_c < \infty$

$$(4.33b) \quad m_c \|z\|^2 \leq \left\langle z, \left(\frac{\partial \bar{x}(u)}{\partial u} \right) \left(\frac{\partial \bar{x}(u)}{\partial u} \right)^T z \right\rangle \leq M_c \|z\|^2, \quad \forall z \in \mathbb{R}^n, \quad u \in \mathbb{R}^N.$$

Suppose that $M_c M_g \geq 2$, that \hat{x} is the unique minimizer of $g(\cdot)$, and that $\{u_i\}_{i=0}^\infty$ is a sequence constructed by the Armijo method in solving problem (4.33a). Then

$$(4.34a) \quad (i) \quad \lim_{i \rightarrow \infty} \nabla g(\bar{x}(u_i)) = 0,$$

$$(4.34b) \quad (ii) \quad \lim_{i \rightarrow \infty} c(u_i) = g(\hat{x}),$$

$$(4.34c) \quad (iii) \quad \overline{\lim}_{i \rightarrow \infty} \frac{c(u_{i+1}) - g(\hat{x})}{c(u_i) - g(\hat{x})} \leq 1 - \frac{4(1-\alpha)\alpha\beta m_g m_c}{M_g M_c},$$

- (iv) The sequence $\{\bar{x}(u_i)\}_{i=0}^{\infty}$ converges R -linearly to \hat{x} ,
- (v) There exists a $\hat{u} \in \mathbb{R}^N$ such that $\nabla c(\hat{u}) = 0$ and the sequence $\{\|u_i - \hat{u}\|\}_{i=0}^{\infty}$ converges R -linearly to 0.

Again we see that the use of adaptive discretization is preferable to fixed discretization.

5. Numerical results. Our Adaptive Discretization Algorithm 2.1 and the corresponding version using fixed discretization were coded in C and both were executed on a SUN Sparc Workstation. In the experiments below, the algorithm parameters common to both versions, were set as follows: $\alpha = 0.9$, $\beta = 0.9$.

For Algorithm 2.1, we set $q_{-1} = 6$, $\tau = 1.0$, $\sigma = 1.25$, and $D = (10.0^{-6})^{\sigma}(q_{\text{Fix}})^{\tau}$, with $q_{\text{Fix}} = 200$.

The number of discretization points used in the fixed discretization version was $q_{\text{Fix}} = 200$.

To construct a meaningful common stopping criterion for both algorithms, for each test problem we determined a threshold value of the cost that was close to the optimum value, and stopped the fixed discretization version of the algorithm when final function value $\psi_{q_{\text{Fix}}}(x_{\text{final}}) \leq \psi_{\text{threshold}}$. We adopted a more stringent stopping criterion for the Adaptive Discretization Algorithm 2.1, which was stopped when the final function value $\psi_{q_{\text{final}}}(x_{\text{final}}) \leq \psi_{\text{threshold}}$ and $q_{\text{final}} \geq q_{\text{Fix}}$.

The following three test problems were selected from [17].

Problem 1 (see [17]). The following holds:

$$(5.1) \quad \begin{aligned} \psi(x) = & (x_1)^2 + (x_2)^2 + (x_3)^2 \\ & + \max \left\{ \max_{y \in [0,1]} (x_1 + x_2 \exp(x_3 y) + \exp(2y) - 2 \sin(4y)), 0 \right\}; \end{aligned}$$

initial point $x_0 = (1.0, 1.0, 1.0)$, $\psi_{200}(x_0) = 15.6209$, $\psi_{\text{threshold}} = 5.3347$.

Problem 2 (see [17]). The following holds:

$$(5.2) \quad \psi(x) = x_1 + x_2/2 + x_3/3 + \max \left\{ \max_{y \in [0,1]} (\tan(y) - x_1 - x_2 y - x_3 y^2), 0 \right\};$$

initial point $x_0 = (0.0, 0.0, 0.0)$, $\psi_{200}(x_0) = 1.5574$, $\psi_{\text{threshold}} = 0.6491$.

Problem 3 (see [17]). The following holds:

$$(5.3) \quad \begin{aligned} \psi(x) = & x_1 + x_2/2 + x_3/3 + x_4/4 + x_5/5 + x_6/6 \\ & + \max \left\{ \max_{y \in [0,1]} (\tan(y) - x_1 - x_2 y - x_3 y^2 - x_4 y^3 - x_5 y^4 - x_6 y^5), 0 \right\}; \end{aligned}$$

initial point $x_0 = (0.0, 0.0, 0.0, 0.0, 0.0, 0.0)$, $\psi_{200}(x_0) = 1.5574$, $\psi_{\text{threshold}} = 0.61682$.

Table 1 summarizes the performance of the two discretization schemes on the above three test problems. We report the performance of the two discretization schemes in terms of the composite number NT , which is the sum of the total number of function evaluations NF and the total number of gradient evaluations NG multiplied by the number of variables, i.e., $NT = NF + nNG$. It is clear from Table 1 that Adaptive Discretization Algorithm 2.1 outperformed the fixed discretization versions by a factor ranging from about 1.5 to 4, in terms of the measure NT .

6. Conclusion. There is an accumulation of empirical evidence to support the claim that, in skillful hands, adaptive discretization schemes can produce considerable computational savings in the solution of optimization problems that must be discretized.

TABLE 1
Summary of numerical results.

Problems	Algorithm 2.1			Fixed Discretization Algorithm		
	NT	q_{final}	$\psi_{q_{\text{final}}}(x_{\text{final}})$	NT	q_{Fix}	$\psi_{q_{\text{Fix}}}(x_{\text{final}})$
1	8,860	301	5.334687	39,600	200	5.334689
2	11,844	200	0.6490467	26,000	200	0.6490803
3	59,063	301	0.6168118	78,600	200	0.6168186

However, prior to the work presented in this paper, there was no automatic discretization scheme for first-order algorithms, whose computational savings could be predicted on the basis of analysis and whose overall rate of convergence could be established. We expect that the discretization techniques presented in this paper will prove to be of practical importance in engineering design and optimal control.

Acknowledgments. We thank the referees for their most helpful comments.

REFERENCES

- [1] L. ARMIJO, *Minimization of functions having continuous partial derivatives*, Pacific J. Math., 16 (1966), pp. 1-3.
- [2] B. M. ALEKSEEV, V. M. TIKHOMIROV, AND S. V. FOMIN, *Optimalnoye Upravleniye (Optimal Control)*, Nauka, Moscow, 1979.
- [3] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.
- [4] J. W. DANIEL, *The Approximation Minimization of Functionals*, Prentice-Hall, Englewood Cliffs, NJ, 1971.
- [5] J. C. DUNN, *Diagonally modified conditional gradient methods for input constrained optimal control problems*, SIAM J. Control Optim., 24 (1986), pp. 1177-1191.
- [6] L. HE AND E. POLAK, *An optimal diagonalization strategy for the solution of a class of optimal design problems*, IEEE Trans. Automat. Control, 35 (1990), pp. 258-267; Electronics Research Laboratory, University of California, Berkeley, CA, Memorandum No. UCB/ERL M88/41, 1988.
- [7] R. KLESSIG AND E. POLAK, *An adaptive precision gradient method for optimal control*, SIAM J. Control, 11 (1973), pp. 80-93.
- [8] D. G. LUENBERGER, *Introduction to Linear and Nonlinear Systems*, Addison-Wesley, New York, 1983.
- [9] O. PIRONNEAU AND E. POLAK, *On the rate of convergence of certain methods of centers*, Math. Programming, 2 (1972), pp. 230-258.
- [10] E. POLAK, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, 1971.
- [11] ———, *On the mathematical foundations of nondifferentiable optimization in engineering design*, SIAM Rev., 29 (1987), pp. 21-91.
- [12] ———, *Basics of minimax algorithms*, in Nonsmooth Optimization and Related Topics, F. H. Clarke, V. F. Dem'yanov, and F. Giannessi, eds., Plenum Press, New York, 1989, pp. 343-367.
- [13] E. POLAK AND L. HE, *A unified phase I-phase II method of feasible directions for semi-infinite optimization*, Electronics Research Laboratory, University of California, Berkeley, CA, Memorandum No. UCB/ERL No. M89/7, Feb. 1989; J. Optim. Theory Appl., to appear.
- [14] E. POLAK AND E. J. WIEST, *Domain rescaling techniques for the solution of affinely parametrized nondifferentiable optimal design problems*, in Proc. 27th IEEE Conference on Decision and Control, Austin, TX, Dec. 7-9, 1988.
- [15] B. N. PSHENICHNYI AND YU. M. DANILIN, *Numerical Methods in Extremal Problems (Chislennyye Metody v Ekstremal'nykh Zadachakh)*, Nauka, Moscow, 1975.
- [16] S. S. SASTRY AND C. A. DESOER, *The robustness of controllability and observability of linear time-varying systems*, IEEE Trans. Automat. Control, 27 (1982), pp. 933-939.

- [17] Y. TANAKA, M. FUKUSHIMA, AND T. IBARAKI, *A comparative study of several semi-infinite nonlinear programming algorithms*, European J. Oper. Res., 36 (1988), pp. 92–100.
- [18] C. T. KELLEY AND J. L. NORTHRUP, *A pointwise quasi-Newton method for integral equations*
- [19] E. SACHS, *Rates of convergence for adaptive Newton methods*, J. Optim. Theory Appl., 48 (1986), pp. 175–190.
- [20] S. LANG, *Real Analysis*, 2nd ed. Addison-Wesley, Reading, MA, 1983.

ON ROBUST PI-CONTROL OF INFINITE-DIMENSIONAL SYSTEMS*

HARTMUT LOGEMANN† AND HANS ZWART‡

Abstract. A PI-controller is applied to a class of linear multivariable infinite-dimensional minimum-phase systems satisfying a generalized “relative-degree one” condition. It is shown that the closed-loop system is stable and tracks asymptotically constant reference signals in the presence of asymptotically constant disturbances, provided that the controller gains are sufficiently large. It turns out that the closed-loop system has nice robustness properties under high-gain conditions. In particular, robustness criteria for external and internal stability are given if the closed-loop system is subjected to perturbations induced by nonlinearities in the feedback loop. The analysis is based on frequency-domain as well as state-space methods.

Key words. infinite-dimensional systems, PI-controllers, high-gain feedback control, robust stability, robust tracking, measurement nonlinearities, input-output stability, internal stability

AMS(MOS) subject classifications. 93C25, 93C35, 93D15, 93D20, 93D25

1. Introduction. The concepts of classical control theory such as root locus plots, Nyquist diagrams, and PI- and PID-controllers are still very popular among control engineers because of their simplicity and their applicability to a great variety of practical problems. Designs based on classical frequency-domain methods lead to low-dimensional controllers, which are easy to implement. Although it was developed mainly for finite-dimensional systems, classical control theory has been applied by engineers to infinite-dimensional systems for many years, despite the fact that few precise theoretical results were available. Since the late 1970s, there has been a renewed theoretical interest in the use of methods from classical control theory for designing control laws for (multivariable) infinite-dimensional systems; see, e.g., Pohjolainen [33], Banks and Abbasi-Ghelnansarai [1], and Byrnes and Gilliam [3] for root-locus techniques; Boyd and Desoer [2] and Freudenberg and Looze [9] for a priori performance bounds on feedback systems such as Bode-type integral relationships; Desoer and Wang [7], Harris and Valenca [11], and Logemann [18], [19], [21] for Nyquist-type stability criteria; Pohjolainen [34], Jussila and Koivo [15], Kobayashi [17], and Logemann and Owens [24] for low-gain PI-control; and Logemann and Owens [22], [23] for high-gain PI-control.

In this paper we continue the work on high-gain PI-control of infinite-dimensional systems started by Logemann and Owens [22], [23]. We investigate stability, tracking, disturbance rejection, and robustness properties achieved by a high-gain PI-controller applied to an infinite-dimensional minimum-phase system satisfying a generalized “relative-degree one” condition. In particular, we study the robustness of closed-loop stability with respect to various classes of measurement nonlinearities. Our analysis is based on time-domain input-output methods, frequency-domain methods, as well as state-space methods. To relate frequency-domain results, on one hand, and state-space results, on the other hand, we express frequency-domain conditions in state-space terms, and vice versa (cf. § 4). Moreover, recent results on the relationship between input-output stability and internal stability of linear infinite-dimensional systems (see Jacobson [13], Jacobson and Nett [14], and Curtain [5]) will play an important role in § 5.

* Received by the editors May 14, 1990; accepted for publication (in revised form) December 18, 1990.

† University of Bremen, Institute for Dynamical Systems, Postfach 330 440, 2800 Bremen 33, Germany.

‡ University of Twente, Department of Applied Mathematics, Post Office Box 217, 7500 AE Enschede, the Netherlands.

The content of the paper is as follows. Section 2 contains some preliminaries and introduces the notation used in the sequel. In § 3 we consider systems described by (not necessarily rational) transfer matrices G of the form

$$(1.1) \quad G(s) = \left(I + \frac{1}{s} D^{-1} H(s) \right)^{-1} \frac{1}{s} D^{-1},$$

where H is a “stable” transfer matrix (see § 3 for a precise definition) and D is a nonsingular constant matrix. We derive a necessary and sufficient condition for a transfer matrix to be of the form (1.1) in terms of its zeros and its behaviour as $|s| \rightarrow \infty$ in the right half-plane. A PI-controller is given that achieves input-output stability and tracking of asymptotically constant reference signals in the presence of asymptotically constant disturbances, provided that the controller gains are sufficiently large. It turns out that the transient performance of the closed-loop system improves as the controller gains increase and that it is perfect for infinitely large gains. Moreover, we investigate the robustness of closed-loop stability with respect to (possibly time-varying) finite-gain stable nonlinearities in the feedback loop. Static, as well as dynamic, nonlinearities are considered, and sufficient conditions for input-output stability of the nonlinear closed-loop system are given. We emphasize that wellposedness questions (i.e., the problem of existence and uniqueness of solutions) are carefully treated. Sections 4 and 5 are devoted to the special situation when the plant transfer matrix G can be realized by an abstract infinite-dimensional state-space system with bounded control and observation operators. In § 4 we prove that the zeros of an infinite-dimensional state-space system (as defined, e.g., in Zwart [40]) coincide with the zeros of its transfer matrix (as defined in § 2), provided that the system is exponentially stabilizable and exponentially detectable. Furthermore, we derive various sufficient conditions in state-space terms for (1.1) to be satisfied. In § 5 we deal with the problem of *internal* stability of the closed-loop system perturbed by nonlinearities in the feedback loop. Assuming that the realizations of the plant and the controller are both exponentially stabilizable and exponentially detectable, we show that the criterion for input-output stability given in § 3 also ensures global exponential stability of the nonlinear feedback scheme if the nonlinearities in the loop are static. In the case of dynamical nonlinearities, we can prove that the origin of the closed-loop system is globally attractive. The proofs of some of the results in §§ 3–5 are relegated to Appendices 1 and 2. For the convenience of the reader, we have included some recent material on exponential stabilizability and exponential detectability of infinite-dimensional systems in Appendix 3.

2. Notation and preliminaries.

- $\mathbb{C}_\alpha := \{s \in \mathbb{C} \mid \operatorname{Re}(s) > \alpha\}$, $\alpha \in \mathbb{R}$.
- Let $U \subset \mathbb{C}$ be open, then $\mathcal{H}(U)$ and $\mathcal{M}(U)$ denote the holomorphic and meromorphic functions on U , respectively.
- $H_\alpha^\infty := \{f: \mathbb{C}_\alpha \rightarrow \mathbb{C} \mid f \text{ bounded and holomorphic}\}$.
- $H_-^\infty := \bigcup_{\alpha < 0} H_\alpha^\infty$.
- Consider distributions of the form

$$(2.1) \quad f = f_a + \sum_{j=0}^{\infty} f_j \delta_{t_j},$$

where $f_a: \mathbb{R}_+ \rightarrow \mathbb{C}$ is measurable, $f_j \in \mathbb{C}$, $t_0 = 0$, $t_j > 0$ for $j \geq 1$, and δ_{t_j} denotes the Dirac distribution with support in $\{t_j\}$. Let \mathcal{A} be the set of all distributions f of the form (2.1) such that

$$(2.2) \quad \|f\|_{\mathcal{A}} := \int_0^\infty |f_a(t)| dt + \sum_{j=0}^\infty |f_j| < \infty.$$

\mathcal{A} is a convolution algebra and, provided with the norm given by (2.2), it becomes a Banach algebra (cf. Hille and Phillips [12, p. 141]).

- $\mathcal{A}_- := \{f \in \mathcal{A} \mid \text{there exists } \varepsilon > 0: fe^{\varepsilon \cdot} \in \mathcal{A}\}$.
- $\hat{\mathcal{A}}, \hat{\mathcal{A}}_-$ is the set consisting of the Laplace transformed elements of $\mathcal{A}, \mathcal{A}_-$, respectively. Realize that $\hat{\mathcal{A}} \subset H_0^\infty$ and $\hat{\mathcal{A}}_- \subset H_-^\infty$.
- \check{f} denotes the inverse Laplace transform of f .
- θ denotes the unit step.
- Let $M = (m_{ij}) \in \mathbb{C}^{p \times p}$, then $\|M\| := \max_{1 \leq i \leq p} \sum_{j=1}^p |m_{ij}|$ unless stated otherwise, $\bar{\sigma}(M) :=$ largest singular value of M , $W(M) :=$ numerical range of M (cf. Halmos [10]).
- Let X be a Banach space and $A: D(A) \subset X \rightarrow X$ a linear operator, then $\sigma(A) :=$ spectrum of A , $\sigma_p(A) :=$ point spectrum of A , and $\rho(A) :=$ resolvent set of A .
- For $F = (f_{ij}) \in (L^1(\mathbb{R}_+))^{p \times p}$ define $\|F\|_1 := \max_{1 \leq i \leq p} \sum_{j=1}^p \|f_{ij}\|_1$.
- Let $F = (f_{ij}) \in \mathcal{A}^{p \times p}$, then $\|F\|_{\mathcal{A}} := \max_{1 \leq i \leq p} \sum_{j=1}^p \|f_{ij}\|_{\mathcal{A}}$.
- If $f = (f_1, \dots, f_p)^t \in (L^q(\mathbb{R}_+))^p$, then $\|f\|_q := \max_{1 \leq j \leq p} \|f_j\|_q$.
- For $F \in (H_0^\infty)^{p \times p}$ define $\|F\|_\infty := \sup_{s \in \mathbb{C}_0} \bar{\sigma}(F(s))$.
- Let f be a function defined on an interval $[a, b)$, $a < b \leq \infty$; then we define, for $t \geq a$,

$$(\pi_t f)(\tau) := \begin{cases} f(\tau), & a \leq \tau \leq t, \\ 0, & \tau > t. \end{cases}$$

- Define the space $LL^q(\mathbb{R}_+)$ by $LL^q(\mathbb{R}_+) := \{f: \mathbb{R}_+ \rightarrow \mathbb{C} \mid f \text{ measurable and } \pi_t f \in L^q(\mathbb{R}_+) \text{ for all } t \geq 0\}$, i.e., $f \in LL^q(\mathbb{R}_+)$, if and only if $|f|^q$ is locally integrable.

We need three additional concepts:

H_-^∞ -stability. Let $G \in \mathcal{M}(\mathbb{C}_\alpha)^{p \times p}$ and $K \in \mathcal{M}(\mathbb{C}_\alpha)^{q \times p}$ for some $\alpha < 0$; then it is convenient to denote the feedback system shown in Fig. 1 by $\mathcal{F}[G, K]$, and we say that $\mathcal{F}[G, K]$ is H_-^∞ -stable if

$$\begin{pmatrix} (I + KG)^{-1}K & -(I + KG)^{-1}KG \\ (I + GK)^{-1}GK & (I + GK)^{-1}G \end{pmatrix}$$

is in $(H_-^\infty)^{(p+q) \times (p+q)}$.

Zeros of a square meromorphic matrix. Let $M \in \mathcal{M}(U)^{p \times p}$. Since it is well known that $\mathcal{M}(U)$ is the quotient field of $\mathcal{H}(U)$ and $\mathcal{H}(U)$ is a Bezout domain¹ (see, e.g., Rudin [35]), it follows that M admits a right coprime factorization over $\mathcal{H}(U)$, i.e., there exist matrices $N, D, X, Y \in \mathcal{H}(U)^{p \times p}$ such that $\det(D) \neq 0$, $M = ND^{-1}$, and $XD + YN = I$. Right coprime factorizations are unique up to multiplication from the right by units of $\mathcal{H}(U)^{p \times p}$ (see, for example, Vidyasagar, Schneider, and Francis [37]). The zeros of M are, by definition, the zeros of $\det(N)$.

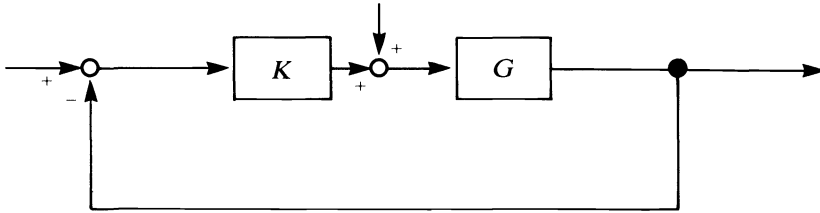


FIG. 1

¹ An integral domain is called Bezout domain if every finitely generated ideal is principal.

Asymptotically constant signals. A function $f: \mathbb{R}_+ \rightarrow \mathbb{C}^p$ is called asymptotically constant if there exists $c \in \mathbb{C}^p$ such that $\lim_{t \rightarrow \infty} f(t) = c$.

3. Robust PI-control of infinite-dimensional systems: Results on input-output stability. Let G be a meromorphic transfer function of size $p \times p$ such that on \mathbb{C}_α (for some $\alpha < 0$)

$$(3.1) \quad G^{-1}(s) = sD + H(s), \quad \text{where } D \in \mathbb{C}^{p \times p}, \quad \det(D) \neq 0, \quad \text{and } H \in (H_-^\infty)^{p \times p}.$$

Of course, (3.1) is equivalent to

$$(3.2) \quad G(s) = \left(I + \frac{1}{s} D^{-1} H(s) \right)^{-1} \frac{1}{s} D^{-1}.$$

Hence (3.1) means that G can be decomposed, as shown in Fig. 2. The following proposition gives a necessary and sufficient condition for G to be of the form (3.2).

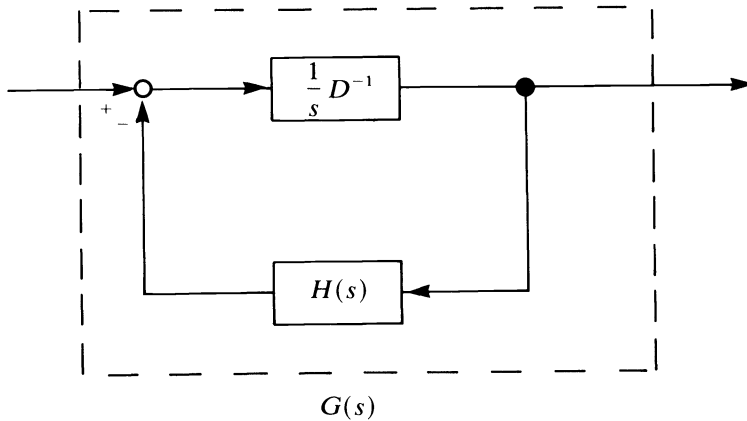


FIG. 2

PROPOSITION 3.1. Let G be a transfer matrix of size $p \times p$, which is meromorphic in \mathbb{C}_α for some $\alpha < 0$. Then G^{-1} is of the form (3.1) if and only if there exist a number β with $\alpha < \beta < 0$ and an invertible matrix $D \in \mathbb{C}^{p \times p}$ such that G has no zeros in $\bar{\mathbb{C}}_\beta$ and $sG(s) - D^{-1} = O(1/s)$ as $|s| \rightarrow \infty$ in \mathbb{C}_β .

As a consequence, we have the following corollary.

COROLLARY 3.2. Suppose that $A \in (H_-^\infty)^{n \times n}$, $B \in \mathbb{C}^{n \times p}$, and $C \in \mathbb{C}^{p \times n}$, and define

$$G(s) := C(sI - A(s))^{-1}B \quad \text{and} \quad \chi(s) := \det \begin{pmatrix} sI - A(s) & -B \\ C & 0 \end{pmatrix}.$$

If $\det(CB) \neq 0$ and if χ has no zeros in $\bar{\mathbb{C}}_0$, then G^{-1} is of the form (3.1) with $D = (CB)^{-1}$.

The proof of the proposition and the corollary can be found in Appendix 1.

We give two classes of infinite-dimensional systems whose transfer matrices satisfy (3.1).

Example 3.3 (Retarded systems). Consider the retarded system

$$\dot{x}(t) = \int_{-h}^0 dA(\tau)x(t+\tau) + Bu(t), \quad y(t) = Cx(t),$$

where $h > 0$ is the length of the delay, the function $A: [-h, 0] \rightarrow \mathbb{R}^{n \times n}$ is of bounded variation, $B \in \mathbb{R}^{n \times p}$, and $C \in \mathbb{R}^{p \times n}$. It is straightforward to show that $\hat{A}(s) := \int_{-h}^0 dA(\tau) e^{s\tau} d\tau$ is holomorphic and bounded on \mathbb{C}_α for any $\alpha \in \mathbb{R}$. In particular, we

have $\hat{A} \in (H_-^\infty)^{n \times n}$. The transfer matrix of the above retarded system is given by $G(s) = C(sI - \hat{A}(s))^{-1}B$. It follows from Corollary 3.2 that G^{-1} is of the form (3.1), provided the conditions $\det(CB) \neq 0$ and

$$\det \begin{pmatrix} sI - \hat{A}(s) & -B \\ C & 0 \end{pmatrix} \neq 0 \quad \text{for all } s \in \bar{\mathbb{C}}_0$$

are satisfied.

Example 3.4 (Volterra integrodifferential systems). Consider the system

$$\dot{x}(t) = A_0 x(t) + \int_0^t A_1(t-\tau)x(\tau) d\tau + Bu(t), \quad y(t) = Cx(t),$$

where $A_0 \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times p}$, $C \in \mathbb{R}^{p \times n}$, and $e^{\varepsilon \cdot} A_1(\cdot) \in L^1(\mathbb{R}_+)$ for some $\varepsilon > 0$. Noting that the Laplace transform \hat{A}_1 of A_1 is in $(H_-^\infty)^{n \times n}$, it follows from Corollary 3.2 that the transfer matrix $G(s) = C(sI - A_0 - \hat{A}_1(s))^{-1}B$ will satisfy (3.1) if $\det(CB) \neq 0$ and

$$\det \begin{pmatrix} sI - A_0 - \hat{A}_1(s) & -B \\ C & 0 \end{pmatrix} \neq 0 \quad \text{for all } s \in \bar{\mathbb{C}}_0.$$

It is fairly obvious that Examples 3.3 and 3.4 can be extended to certain classes of retarded systems with infinite delay and Volterra-Stieltjes integrodifferential systems.

Consider the PI-controller

$$(3.3) \quad K_k(s) := \Gamma \operatorname{diag}_{1 \leq j \leq p} \left(k_j + c_j + \frac{k_j c_j}{s} \right),$$

where $\Gamma \in \mathbb{C}^{p \times p}$, $\det(\Gamma) \neq 0$, $k = (k_1, \dots, k_p)'$, $k_j > 0$, $c_j > 0$ for all $j = 1, 2, \dots, p$. Sometimes it will be useful to emphasize the dependence of the controller (3.3) on the “gain vector” k , since we will be interested in the high-gain situation, where $k_j \rightarrow \infty$, $j = 1, \dots, p$. That is why we introduced the subscript k in (3.3).

The above controller was investigated in Owens and Chotai [31] when applied to finite-dimensional systems. The infinite-dimensional case is studied in Logemann and Owens [22], [23].

The following theorem gives sufficient conditions for robust stability when the controller (3.3) is applied to a system satisfying (3.1).

THEOREM 3.5. *Let G be a transfer matrix such that (3.1) is satisfied. Then (i) the feedback system $\mathcal{F}[G, K_k]$ is H_-^∞ -stable for all sufficiently large k_j , $j = 1, \dots, p$, if*

$$(3.4) \quad \|\Gamma^{-1}(\Gamma - D)\| < 1,$$

where $\|\cdot\|$ is any submultiplicative norm on $\mathbb{C}^{p \times p}$ with the additional property that $\|\operatorname{diag}(a_j)\| \leq \max_j |a_j|$ for arbitrary $a_1, \dots, a_p \in \mathbb{C}$; and (ii) under the additional assumption that $k_j = \gamma_j \kappa$ with $\gamma_j > 0$ fixed ($j = 1, \dots, p$) the feedback scheme $\mathcal{F}[G, K_k]$ is H_-^∞ -stable for a sufficiently large κ if

$$(3.5) \quad \sigma \left(\operatorname{diag}_j(\gamma_j) \Gamma^{-1} D \right) \subset \mathbb{C}_0$$

or

$$(3.6) \quad W(\Gamma^{-1} D) \subset \mathbb{C}_0.$$

Proof. See Logemann and Owens [22].

Remark 3.6. (i) Note that K_k does not depend on H . Trivially, (3.4)–(3.6) are satisfied if $\Gamma = D$. Obviously, $\Gamma = D$ would be a natural choice in (3.3). However, D might not be exactly known to the designer.

(ii) Further applications of the concept of numerical range to control problems can be found in Mees [25].

To study the tracking and output disturbance rejection properties of $\mathcal{F}[G, K_k]$ (cf. Fig. 3), define $L_k := (I + GK_k)^{-1}GK_k$ and $H_k := (I + GK_k)^{-1}$.

PROPOSITION 3.7. *Let G be a square meromorphic transfer matrix such that G^{-1} is of the form (3.1). If $\mathcal{F}[G, K_k]$ is H^∞ -stable, then the closed-loop system tracks asymptotically constant reference signals in the presence of asymptotically constant output disturbances, i.e., for $r: \mathbb{R}_+ \rightarrow \mathbb{C}^p$ and $d: \mathbb{R}_+ \rightarrow \mathbb{C}^p$ such that $\lim_{t \rightarrow \infty} r(t) = r_\infty$ and $\lim_{t \rightarrow \infty} d(t) = d_\infty$, we have*

$$\lim_{t \rightarrow \infty} (\check{L}_k * r)(t) = r_\infty \quad \text{and} \quad \lim_{t \rightarrow \infty} (\check{H}_k * d)(t) = 0.$$

Proof. It follows from the stability of $\mathcal{F}[G, K_k]$ that $L_k, H_k \in (H^\infty)^{p \times p}$. Now it is easy to see that $sL_k(s)$ and $s(H_k(s) - I)$ are bounded on \mathbb{C}_β for some $\beta < 0$, and hence we obtain, using a result in Mossaheb [27] (cf. also Logemann [18]), that $e^{\alpha \cdot} \check{L}_k \in (L^1(\mathbb{R}_+))^{p \times p}$ and $e^{\alpha \cdot} \check{H}_k \in (\mathbb{R}\delta_0 + L^1(\mathbb{R}_1))^{p \times p}$ for all $\alpha \in (0, -\beta)$, and thus L_k and $H_k \in (\hat{\mathcal{A}}_-)^{p \times p}$. By the final value theorem for transfer functions in $\hat{\mathcal{A}}_-$ (cf. Callier and Winkin [4]), it is sufficient to show that $L_k(0) = I$ and $H_k(0) = 0$.

An elementary calculation gives

$$\begin{aligned} L_k(s) &= ((G(s)K_k(s))^{-1} + I)^{-1} \\ &= \left\{ \text{diag} \left(\frac{s}{s(k_j + e_j) + k_j c_j} \right) \Gamma^{-1}(sD + H(s)) + I \right\}^{-1}. \end{aligned}$$

Since $k_j, c_j > 0$, $j = 1, \dots, p$, it follows that $L_k(0) = I$. Since $H_k = I - L_k$, we obtain $H_k(0) = 0$. \square

To investigate the transient performance of $\mathcal{F}[G, K_k]$ and the robustness of closed-loop stability with respect to measurement nonlinearities, the following lemma is useful.

LEMMA 3.8. *Let G be a square meromorphic transfer matrix such that G^{-1} is of the form (3.1), define K_k as in (3.3), and set*

$$G^*(s) := \frac{1}{s} \Gamma^{-1} \quad \text{and} \quad L_k^*(s) := (I + G^* K_k)^{-1} G^* K_k.$$

Then the following hold: (i) If $\Gamma = D$, we have

$$(3.7) \quad \lim_{k \rightarrow \infty} \|\check{L}_k^* - \check{L}_k\|_1 = 0$$

(by $k \rightarrow \infty$, we mean $\min_{1 \leq j \leq p} (k_j) \rightarrow \infty$);

(ii) If $\|\Gamma^{-1}(\Gamma - D)\| =: \varepsilon < \frac{1}{2}$, then

$$(3.8) \quad \limsup_{k \rightarrow \infty} \|\check{L}_k^* - \check{L}_k\|_1 \leq \frac{2\varepsilon}{1 - 2\varepsilon}.$$

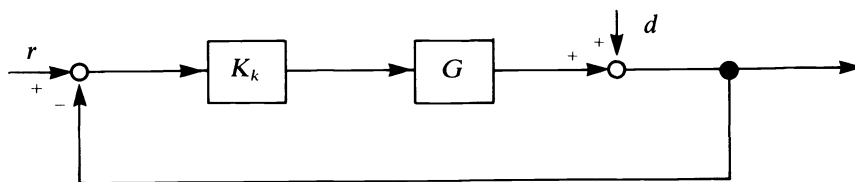


FIG. 3

The proof of Lemma 3.8 can be found in Appendix 2. Part (i) was proved in Logemann and Owens [22] under the extra assumption that H in (3.1) belongs to $\hat{\mathcal{A}}_-^{p \times p}$. It should be noted that it is, in general, considerably more difficult to check if a given transfer function is in $\hat{\mathcal{A}}_-$ than to verify that it belongs to H_-^∞ .

Remark 3.9. (i) Define $\theta_j := e_j \theta$, where $e_j = (0, \dots, 1, \dots, 0)^t \in \mathbb{R}^p$. We consider the transient performance of the feedback system $\mathcal{F}[G^*, K_k]$. It is easy to see that the transfer matrix L_k^* is given by

$$L_k^*(s) = \text{diag} \left(\frac{(k_j + c_j)s + k_j c_j}{(s + k_j)(s + c_j)} \right)_{1 \leq j \leq p}.$$

Since we are interested in high-gain feedback, we may assume without loss of generality that $k_j > c_j$, $j = 1, \dots, p$. A routine calculation gives the following estimates for the overshoot O_j , the rise time T_j^r , and the settling time T_j^s in the j th loop (see, e.g., Franklin, Powell, and Emami-Naeini [8] for the notions of overshoot, rise time, and so forth).

We have $O_j \leq c_j / (k_j - c_j)$, $T_j^r \leq \frac{1}{2} \tau_j$, and $T_j^s \leq \max(\frac{1}{2} \tau_j, -1/c_j \ln((k_j - c_j)/100 c_j))$, where $\tau_j = (2(\ln(k_j) - \ln(c_j)) / (k_j - c_j))$ is the time when the maximal overshoot occurs. The settling time T_j^s is defined here as the time required for the signal $(L_k^* * \theta_j)_j(t)$ to stay within the interval $[0.99, 1.01]$.

The estimates show that the transient performance of the feedback system $\mathcal{F}[G^*, K_k]$ improves as the gains k_j , $j = 1, \dots, p$, increase.

(ii) Suppose that $\Gamma = D$. In this case, part (i) of this remark and (3.7) show that the transient performance of the feedback system $\mathcal{F}[G, K_k]$ improves as the gains k_j , $j = 1, \dots, p$ increase. If $\Gamma \neq D$, then (3.8) gives a bound on the performance degradation, provided that the condition of Lemma 3.8 (ii) is satisfied.

In the following, we investigate the effect of measurement nonlinearities on the stability of $\mathcal{F}[G, K_k]$. First, we will concentrate on memoryless nonlinearities. We consider functions $\varphi: \mathbb{R}_+ \times \mathbb{C}^p \rightarrow \mathbb{C}^p$, which satisfy the following conditions:

- (N1) $\varphi(t, x)$ is continuous in t and locally Lipschitz continuous in x , uniformly in t on bounded intervals;
- (N2) φ is unbiased, i.e., $\varphi(t, 0) = 0$, for all $t \geq 0$;
- (N3) $\varphi = \text{id}_{\mathbb{C}^p} + \varphi_1 + \varphi_2$, where φ_1 and φ_2 satisfy

$$\text{and } \left. \begin{array}{l} |\varphi_1(t, x)| \leq \lambda_1 |x| \\ |\varphi_2(t, x)| \leq \lambda_2 \end{array} \right\} \text{ for all } t \geq 0, x \in \mathbb{C}^p, \text{ and some constants } \lambda_1, \lambda_2 \geq 0.$$

Furthermore, for a function $\varphi: \mathbb{R}_+ \times \mathbb{C}_p \rightarrow \mathbb{C}^p$, let N_φ denote the operator induced by φ , i.e., $(N_\varphi u)(t) = \varphi(t, u(t))$ for any function $u: \mathbb{R}_+ \rightarrow \mathbb{C}^p$.

The following result gives a sufficient condition for the stability of the feedback system shown in Fig. 4.

THEOREM 3.10. *Let G be a square transfer matrix such that G^{-1} is of the form (3.1). Let the controller K_k be given by (3.3), suppose that $\varphi: \mathbb{R}_+ \times \mathbb{C}^p \rightarrow \mathbb{C}^p$ satisfies*

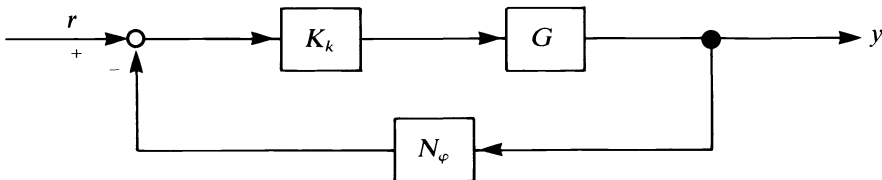


FIG. 4

(N1)–(N3), and assume that the reference signal r is bounded on bounded intervals. Then, if $\lambda_1 < 1$ and

$$(3.9) \quad \|\Gamma^{-1}(\Gamma - D)\| < \frac{1}{2}(1 - \lambda_1),$$

there exists $k^* > 0$ such that for all $k_j > k^*$ ($j = 1, \dots, p$) the nonlinear feedback system given by

$$(3.10) \quad y = \check{G} * e, \quad e = \check{K}_k * (r - N_\varphi(y))$$

is well posed (i.e., there exists a unique globally defined solution of (3.10)) and L^∞ -stable in the sense that there exist nonnegative constants m_1 and m_2 such that

$$(3.11) \quad \|y\|_\infty \leq m_1 + m_2 \|r\|_\infty$$

for all $r \in (L^\infty(\mathbb{R}_+))^p$.

Proof. First, realize that for all sufficiently large k_j ($j = 1, \dots, p$)

$$(3.12) \quad \|\check{L}_k\|_1 \lambda_1 < 1.$$

This follows from (3.9), Lemma 3.8(ii), and the fact that $\lim_{k \rightarrow \infty} \|\check{L}_k^*\|_1 = 1$ (cf. Logemann and Owens [22]), where \check{L}_k^* is defined as in Lemma 3.8. Equation (3.10) can be written in the form

$$(3.13) \quad y = \check{L}_k * r - \check{L}_k * (N_{\varphi_1}(y) + N_{\varphi_2}(y)),$$

which is a nonlinear Volterra integral equation in y . It follows from Theorem 1.2 and Corollary 2.7, in Miller [26] that (3.13) has a unique (continuous) solution that can be extended to the right as long as it remains bounded. Now pick $t > 0$ such that the solution of (3.13) exists on $[0, t]$. Application of the truncation operator π_t to (3.13) gives

$$(3.14) \quad \begin{aligned} \|\pi_t(y - y_l)\|_\infty &\leq \|\pi_t \check{L}_k * (\pi_t(N_{\varphi_1}(y) + N_{\varphi_2}(y)))\|_\infty \\ &\leq \|\check{L}_k\|_1 (\|\pi_t(N_{\varphi_1}(y))\|_\infty + \lambda_2) \\ &\leq \|\check{L}_k\|_1 (\lambda_1 \|\pi_t y\|_\infty + \lambda_2), \end{aligned}$$

where $y_l := \check{L}_k * r$ is the output of the linear feedback system $\mathcal{F}[G, K_k]$. Setting $\lambda := \|\check{L}_k\|_1$, we obtain

$$(1 - \lambda \lambda_1) \|\pi_t(y - y_l)\|_\infty \leq \lambda (\lambda_1 \|\pi_t y\|_\infty + \lambda_2),$$

and hence ($\lambda \lambda_1 < 1$ by (3.12))

$$(3.15) \quad \|\pi_t(y - y_l)\|_\infty \leq \frac{\lambda}{1 - \lambda \lambda_1} (\lambda_1 \|\pi_t y\|_\infty + \lambda_2).$$

Inequality (3.15) shows that the solution of (3.13) exists on \mathbb{R}_+ , since $\|\pi_t y_l\|_\infty$ is finite for all $t \in \mathbb{R}_+$. Moreover, it follows that (3.11) holds with $m_1 = \lambda \lambda_2 (1 - \lambda \lambda_1)^{-1}$ and $m_2 = \lambda (1 - \lambda \lambda_1)^{-1}$. \square

Remark 3.11. (i) The proof shows that Theorem 3.10 remains true if we replace ∞ by $q = 1, 2, 3, \dots$, provided that $\varphi_2 \equiv 0$.

(ii) Equation (3.15) yields an upper bound on the difference of the output signals of the linear and nonlinear feedback system corresponding to the same input signal r .

We now turn our attention to a certain class of dynamical measurement nonlinearities. We consider operators Φ from $(LL^q(\mathbb{R}_+))^p$ onto itself, which satisfy the following assumptions:

- (N4) Φ is causal, i.e., $\pi_t \Phi = \pi_t \Phi \pi_t$ for all $t \geq 0$ (cf. Willems [38]);
- (N5) Φ is locally Lipschitz continuous, i.e., for all $t \geq 0$ there exists $l_t \geq 0$ such that $\|\pi_t(\Phi u - \Phi v)\|_q \leq l_t \|\pi_t(u - v)\|_q$ for all $u, v \in (LL^q(\mathbb{R}_+))^p$ (cf. Willems [38]);
- (N6) Φ is unbiased, i.e., $\Phi(0) = 0$;
- (N7) $\Phi = \text{id} + \Psi$, where Ψ satisfies $\|\Psi u\|_q \leq \lambda_1 \|u\|_q + \lambda_2$ for all $u \in (L^q(\mathbb{R}_+))^p$ and some nonnegative constants λ_1 and λ_2 .

Remark 3.12. Consider the nonlinearity N_φ induced by the function $\varphi: \mathbb{R}_+ \times \mathbb{C}^p \rightarrow \mathbb{C}^p$. Then, in general, N_φ will not fulfill (N5) unless φ satisfies a global Lipschitz condition.

THEOREM 3.13. *Let G be a square transfer matrix such that G^{-1} is of the form (3.1), let the controller K_k be given by (3.3), suppose that $\Phi: (LL^q(\mathbb{R}_+))^p \rightarrow (LL^q(\mathbb{R}_+))^p$ satisfies (N4)–(N7), and assume that $r \in (LL^q(\mathbb{R}_+))^p$. Then we have (i) the nonlinear feedback system given by*

$$(3.16) \quad y = \check{G} * e, \quad e = \check{K}_k * (r - \Phi(y))$$

is well posed in the sense that there exists a unique (globally defined) solution $y \in (LL^q(\mathbb{R}_+))^p$; and (ii) suppose that $\lambda_1 < 1$ and

$$(3.17) \quad \|\Gamma^{-1}(\Gamma - D)\| < \frac{1}{2}(1 - \lambda_1);$$

then for all sufficiently large k_j ($j = 1, \dots, p$), the feedback system (3.16) is L^q -stable in the sense that there exist nonnegative constants m_1 and m_2 such that $\|y\|_q \leq m_1 + m_2 \|r\|_q$, for all $r \in (L^q(\mathbb{R}_+))^p$.

Proof. Equation (3.16) can be written in the form

$$(3.18) \quad y = \check{L}_k * r - \check{L}_k * \Psi(y).$$

It follows from Corollary 4.1.2 in Willems [38] that (3.18) admits a unique solution in $(LL^q(\mathbb{R}_+))^p$. The stability result can be shown, as in the proof of Theorem 3.10. \square

4. Conditions in state-space terms for (3.1) to be satisfied. In this section we will give sufficient conditions for a system in state-space form to satisfy the decomposition (3.1). Our state-space system is given by

$$(4.1a) \quad \dot{x} = Ax + Bu; \quad x(0) = x_0,$$

$$(4.1b) \quad y = Cx,$$

where (i) $A: D(A) \subset X \rightarrow X$, X is a Banach space, generates a strongly continuous semigroup, denoted by $T(t)$; (ii) $B: \mathbb{R}^p \mapsto X$; $(u_1, \dots, u_p)' \mapsto \sum_{i=1}^p b_i u_i$, where $b_i \in X$, $i = 1, \dots, p$; and (iii) $C: X \mapsto \mathbb{R}^p$, $x \mapsto (\langle x, c_1 \rangle, \dots, \langle x, c_p \rangle)'$, where $c_i \in X^*$; $i = 1, \dots, p$.

In § 3 we have seen that the zeros of the system play an essential role in determining if it has a decomposition (3.1). The next definition gives an equivalent definition for zeros of a state-space system.

DEFINITION 4.1. Let $z \in \mathbb{C}$; then z is a zero of system (4.1) if the kernel of the operator $\begin{bmatrix} zI - A & B \\ C & 0 \end{bmatrix}: D(A) \oplus \mathbb{C}^p \mapsto X \oplus \mathbb{C}^p$ is nonzero.

LEMMA 4.2. *If system (4.1) is α -exponentially stabilizable and α -exponentially detectable, then the zeros in \mathbb{C}_α of (4.1) are the same as the zeros of $G(s) := C(sI - A)^{-1}B$ as defined in § 2.*

For the definitions of α -exponentially stabilizability and detectability, see Appendix 3.

Proof. Since the state-space system is α -exponentially stabilizable, there exists a bounded linear operator $F: X \rightarrow \mathbb{R}^p$ such that the semigroup generated by $A + BF$, $T_{BF}(t)$, satisfies $\|T_{BF}(t)\| \leq M e^{\beta t}$, $\beta < \alpha$. With this feedback, we can construct the following right coprime factorization of $G(s)$ over $\mathcal{H}(\mathbb{C}_\alpha)$ (see Jacobson [13] and Nett, Jacobson, and Balas [30]):

$$G = NM^{-1}, \quad \text{where } N(s) = C(sI - A - BF)^{-1}B, \quad M(s) = I + F(sI - A - BF)^{-1}B.$$

Let $u \in \mathbb{C}^p$, $u \neq 0$, satisfy $C(s_0I - A - BF)^{-1}Bu = 0$ for some $s_0 \in \mathbb{C}_\alpha$. Then, by the coprimeness of N and M , $u + F(s_0I - A - BF)^{-1}Bu \neq 0$, and

$$(4.2) \quad \begin{bmatrix} s_0I - A & B \\ C & 0 \end{bmatrix} \begin{bmatrix} (s_0I - A - BF)^{-1}Bu \\ -u - F(s_0I - A - BF)^{-1}Bu \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

On the other hand, if

$$(4.3) \quad \begin{bmatrix} s_0I - A & B \\ C & 0 \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

for some $(x, u) \neq (0, 0) \in D(A) \oplus \mathbb{C}^p$, then

$$(4.4) \quad (s_0I - A - BF)x + B(u + Fx) = 0.$$

Premultiplying (4.4) with $C(s_0I - A - BF)^{-1}$ gives

$$0 = Cx + C(s_0I - A - BF)^{-1}B(u + Fx) = 0 + C(s_0I - A - BF)^{-1}B(u + Fx).$$

So, if $(u + Fx) \neq 0$, then s_0 is a zero of the transfer matrix G . If $u + Fx = 0$, then (4.4) with the invertibility of $(s_0I - A - BF)$ would imply that $x = 0$. Hence also $u = 0$, since $u = u + Fx - Fx$. This is in contradiction with $(x, u) \neq (0, 0)$. \square

For the system under consideration, we will prove that under certain conditions the transfer function can be decomposed in a similar way as in (3.1).

LEMMA 4.3. *Suppose that (A, B) is α -exponentially stabilizable. Assume further that $\det(CB) \neq 0$ and let $\gamma \in \mathbb{C}$. Then*

$$(4.5) \quad G^{-1}(s) = s[CB]^{-1} + (\gamma - s)[CB]^{-1}CA(sI - A - BF)^{-1}B[CB]^{-1}$$

holds on $\mathbb{C}_\alpha \cap \rho(A) \cap \rho(A + BF)$, where $G(s) = C(sI - A)^{-1}B$ and

$$(4.6) \quad F := [CB]^{-1}\{-CA + \gamma C\}.$$

Proof. The feedback F defined by (4.6) is an A -degenerate operator (Kato [16, Chap. IV, § 6]). Since (A, B) is α -exponentially stabilizable, the spectrum of A in \mathbb{C}_α is pure point spectrum with finite multiplicity (see Jacobson and Nett [14] or Appendix 3). Together with the fact that A generates a C_0 -semigroup, this implies that $A + BF$ has pure point spectrum with finite multiplicity in \mathbb{C}_α (see Kato [16, Chap. IV, § 6]). So, with the exception of countably many points, we may calculate $(sI - A - BF)^{-1}$ for $s \in \mathbb{C}_2$. Let $s \in \rho(A + BF) \cap \rho(A)$; then

$$\begin{aligned} CA(sI - A - BF)^{-1}B &= CA(sI - A)^{-1}B + CA(sI - A - BF)^{-1}BF(sI - A)^{-1}B \\ &= CA(sI - A)^{-1}B \\ &\quad + CA(sI - A - BF)^{-1}B[CB]^{-1}(-1)CA(sI - A)^{-1}B \\ &\quad + CA(sI - A - BF)^{-1}B[CB]^{-1}\gamma C(sI - A)^{-1}B \\ &= \left\{ I - \left(1 - \frac{\gamma}{s} \right) CA(sI - A - BF)^{-1}B[CB]^{-1} \right\} CA(sI - A)^{-1}B \\ &\quad + \frac{\gamma}{s} CA(sI - A - BF)^{-1}B, \end{aligned}$$

where in the last equality we have used that $C(sI - A)^{-1}B = 1/s\{CB + CA(sI - A)^{-1}B\}$. So

$$\begin{aligned} \left(1 - \frac{\gamma}{s}\right) CA(sI - A - BF)^{-1}B &= \left\{I - \left(1 - \frac{\gamma}{s}\right) CA(sI - A - BF)^{-1}B[CB]^{-1}\right\} \\ &\quad \times CA(sI - A)^{-1}B \\ \Rightarrow CA(sI - A)^{-1}B[CB]^{-1} &= \left\{I - \left(1 - \frac{\gamma}{s}\right) CA(sI - A - BF)^{-1}B[CB]^{-1}\right\}^{-1} \\ &\quad \times \left(1 - \frac{\gamma}{s}\right) CA(sI - A - BF)^{-1}B[CB]^{-1}. \end{aligned}$$

Using the equality $[I - Q(s)]^{-1}Q(s) = [I - Q(s)]^{-1} - I$ gives

$$CA(sI - A)^{-1}B[CB]^{-1} = \left\{I - \left(1 - \frac{\gamma}{s}\right) CA(sI - A - BF)^{-1}B[CB]^{-1}\right\}^{-1} - I.$$

So

$$\begin{aligned} sG(s) &= sC(sI - A)^{-1}B = CB + CA(sI - A)^{-1}B \\ &= \left\{I - \left(1 - \frac{\gamma}{s}\right) CA(sI - A - BF)^{-1}B[CB]^{-1}\right\}^{-1} [CB]. \end{aligned}$$

Thus

$$G^{-1}(s) = s[CB]^{-1} + (\gamma - s)[CB]^{-1}CA(sI - A - BF)^{-1}B[CB]^{-1}. \quad \square$$

So, the above lemma gives a decomposition similar to (3.1), but we do not know if $H(s) = (\gamma - s)[CB]^{-1}CA(sI - A - BF)^{-1}B[CB]^{-1} \in H_-^{\infty(p \times p)}$. This result will be given in Theorem 4.5. First, we will prove that, with the feedback F defined by (4.3), $(sI - A - BF)$ is invertible in \mathbb{C}_α , provided that $G(\cdot)$ has no zeros there.

LEMMA 4.4. *Suppose that (A, B) is α -exponentially stabilizable and (C, A) is α -exponentially detectable. If system (4.1) has no zeros in \mathbb{C}_α , and $\gamma < \alpha$, then $(sI - A - BF)$ is invertible on \mathbb{C}_α , where F is defined by (4.6).*

Proof. From the proof of Lemma 4.3 we know that $A + BF$ has only point spectrum on \mathbb{C}_α . Let λ be an eigenvalue of $(A + BF)$ in \mathbb{C}_α . Then there exists an $x \in X$, $x \neq 0$ such that

$$(\lambda I - A - BF)x = 0 \Leftrightarrow (\lambda I - A + B[CB]^{-1}CA)x - \gamma B[CB]^{-1}Cx = 0.$$

Premultiplying with C gives $\lambda Cx - CAx + CAx - \gamma Cx = 0$. Hence $\lambda = \gamma$ or $Cx = 0$. The first possibility is excluded by assumption, so suppose that $Cx = 0$. Then we have that

$$\begin{pmatrix} \lambda I - A & B \\ C & 0 \end{pmatrix} \begin{pmatrix} x \\ [CB]^{-1}CAx \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Thus λ is a zero, but this is also excluded. So $\sigma(A + BF) \cap \mathbb{C}_\alpha = \emptyset$. \square

So, if the zeros of the system are in $\mathbb{C} \setminus \mathbb{C}_\alpha$, then $G^{-1}(s)$ exists everywhere on \mathbb{C}_α . We may always write $G^{-1}(s)$ as $G^{-1}(s) = sD + H(s)$, where $H(s)$ has no poles in \mathbb{C}_α . However, this is not enough to ensure that $H \in (H_-^\infty)^{p \times p}$. The next theorem will give sufficient conditions for this to hold.

THEOREM 4.5. *Suppose that (A, B) is exponentially stabilizable, (C, A) is exponentially detectable, system (4.1) has no zeros in \mathbb{C}_α for some $\alpha < 0$, and $\det(CB) \neq 0$. Then*

the transfer matrix $G(s) = C(sI - A)^{-1}B$ has the property that $G^{-1}(s) = s[CB]^{-1} + H(s)$ with $H \in (H_-^\infty)^{p \times p}$, provided that either

- (i) $c_i \in D(A^{*2})$; $i = 1, \dots, p$, or
- (ii) $b_i \in D(A)$, $c_i \in D(A^*)$; $i = 1, \dots, p$, or
- (iii) $b_i \in D(A^2)$; $i = 1, \dots, p$, or
- (iv) A generates an analytic semigroup and $b_i \in D(A)$, $i = 1, \dots, p$, or
- (v) A generates an analytic semigroup and $c_i \in D(A^*)$, $i = 1, \dots, p$.

Furthermore,

$$(4.7) \quad H(s) = (\gamma - s)[CB]^{-1}CA(sI - A - BF)^{-1}B[CB]^{-1}$$

with

$$(4.8) \quad F = [CB]^{-1}\{-CA + \gamma C\},$$

where $\gamma < 0$.

Proof. Note that if (A, B, C) is exponentially stabilizable and detectable, then it is also β -exponentially stabilizable and detectable for some $\beta < 0$. So, without loss of generality, we may assume that the system is α -exponentially stabilizable and detectable and that it has no zeros in \mathbb{C}_α for some negative α . Moreover, let $\gamma < \alpha$. These conditions ensure that on \mathbb{C}_α we have by Lemmas 4.3 and 4.4 that

$$G^{-1}(s) = s[CB]^{-1} + (\gamma - s)[CB]^{-1}CA(sI - A - BF)^{-1}B[CB]^{-1}.$$

So we must show that $H(s) := (\gamma - s)[CB]^{-1}CA(sI - A - BF)^{-1}B[CB]^{-1}$ is analytic and bounded on \mathbb{C}_α .

Assume first that $c_i \in D(A^*)$; $i = 1, \dots, p$. So CA is a bounded operator from X to \mathbb{R}^p , and thus F is bounded. Since (A, B) is exponentially stabilizable and since $A + BF$ has no eigenvalues in \mathbb{C}_α (Lemma 4.4), we have that $T_{BF}(t)$ is exponentially stable (see Appendix 3, Theorem A.6), and hence $\gamma[CB]^{-1}CA(s - A - BF)^{-1}B[CB]^{-1}$ is in $(H_-^\infty)^{p \times p}$. This is, in general, not sufficient to ensure that $H \in (H_-^\infty)^{p \times p}$. We can rewrite the operator $sCA(sI - A - BF)^{-1}B$ as

$$\begin{aligned} & CA(sI - A - BF)^{-1}B \\ &= CA(sI - A - BF)(sI - A - BF)^{-1}B + CA(A + BF)(sI - A - BF)^{-1}B \\ (4.9) \quad &= CAB + CA(A + BF)(s - A - BF)^{-1}B \end{aligned}$$

$$(4.10) \quad = CAB + CA(sI - A - BF)^{-1}(A + BF)B.$$

From (4.9) we see that $H \in (H_-^\infty)^{p \times p}$ if $c_i \in D(A^{*2})$; $i = 1, \dots, p$, and from (4.10) we see that $H \in (H_-^\infty)^{p \times p}$ if $c_i \in D(A^*)$ and $b_i \in D(A)$, $i = 1, \dots, p$. So we have proved that conditions (i) and (ii) imply the desired property. Now we will prove that condition (iii) does the same.

If $b_i \in D(A^2)$, $i = 1, \dots, p$, then let us consider a new realization (A_n, B_n, C_n) of $G(s)$, namely $A_n = A$, $B_n = AB$, and $C_n = CA^{-1}$, where we have assumed that $0 \in \rho(A)$, but this is not essential. If (A, B, C) is exponentially stabilizable and exponentially detectable, then (A_n, B_n, C_n) is, also. This follows easily from Theorem A.5 in Appendix 3, the “dual” version of Theorem A.5, and the definitions of A_n , B_n , and C_n . Since $\text{Im } B_n \subset D(A_n)$ and $\text{Im } C_n^* \subset D(A_n^*)$, we have by part (ii) that

$$(4.11) \quad H_n(s) = (\gamma - s)[C_n B_n]^{-1}C_n A_n(sI - A_n - B_n F_n)^{-1}B_n[C_n B_n]^{-1}$$

is an element of $(H_-^\infty)^{p \times p}$, where $F_n := [C_n B_n]^{-1}\{-C_n A_n + \gamma C_n\}$. It is clear that $F_n = FA^{-1}$, with F given by (4.8), and hence $H = H_n$. So it remains to show that condition (iv) or (v) is sufficient, also.

The feedback law F , as defined in (4.8), is an A -degenerate operator. From Zabczyk [39] it is known that if A generates an analytic semigroup, so does $A + BF$, for an A -degenerate feedback F . So, since any analytic semigroup satisfies the spectrum determined growth condition, and $\sigma(A + BF) \cap \mathbb{C}_\alpha = \emptyset$, there exists $M > 0$ such that

$$\|(sI - A - BF)^{-1}\| \leq \frac{M}{|s - \alpha|}; \quad s \in \mathbb{C}_\alpha.$$

Thus, if $c_i \in D(A^*)$; $i = 1, \dots, p$, then

$$\|H(s)\| \leq \frac{|\gamma - s|}{|s - \alpha|} \|CA\| \|[CB]^{-1}\|^2 \|B\| M; \quad s \in \mathbb{C}_\alpha.$$

On the other hand, if $b_i \in D(A)$, then we can rewrite $H(s)$ as

$$H(s) = (\gamma - s)[CB]^{-1}C(sI - A - ABFA^{-1})^{-1}AB[CB]^{-1}; \quad s \in \mathbb{C}_\alpha,$$

and we have that

$$\|H(s)\| \leq \frac{|\gamma - s|}{|s - \alpha|} \|C\| \|AB\| \|[CB]^{-1}\|^2 \tilde{M}; \quad s \in \mathbb{C}_\alpha$$

for some suitable constant $\tilde{M} > 0$. Thus in both cases $H \in (H_-^\infty)^{p \times p}$. \square

Remark 4.6. Since $G(s)$ and $s[CB]^{-1}$ are independent of γ , we must have that $H(\cdot)$ is independent of γ . In fact, for $\gamma > \alpha$ the zero at γ introduced by the term $s - \gamma$ is cancelled by the pole at γ of $(s - A - BF)^{-1}$ (see the proof of Lemma 4.4).

Remark 4.7. Retarded systems do not satisfy any of the smoothness conditions (i)–(v) in Theorem 4.5. However, by Example 3.3, there is a whole class of retarded systems whose transfer matrices admit a decomposition of the form (3.1). This shows that the conditions of Theorem 4.5 are sufficient but not necessary for (3.1) to be satisfied.

5. Internal stability. The stability results in § 3 are formulated in input-output terms. Suppose that the transfer matrix G of a state-space system of the form (4.1) satisfies condition (3.1) (Theorem 4.5 gives conditions in state-space terms for this to be true). If we apply Theorem 3.10 or Theorem 3.13 to G , can we expect internal stability of the closed-loop system? In the *linear* case (i.e., N_φ in Theorem 3.10 and Φ in Theorem 3.10 are equal to the identity) the answer is yes, provided that the state-space realizations of the plant and the controller are both exponentially stabilizable and exponentially detectable. This follows from recent results on the equivalence of input-output and internal stability for infinite-dimensional systems (cf. Jacobson and Nett [14] and Curtain [5]). In this section we investigate the internal asymptotic behaviour of the *nonlinear* feedback systems considered in § 3.

LEMMA 5.1. *Let $T(t)$ be an exponentially stable, strongly continuous semigroup on the Banach space X , denote the generator of $T(t)$ by A , let $B: \mathbb{R}^p \rightarrow X$ and $C: X \rightarrow \mathbb{R}^p$ be bounded linear operators, and suppose that $f: \mathbb{R}_+ \times \mathbb{R}^p$ satisfies (N1) and $|f(t, x)| \leq \lambda|x|$ for all $t \geq 0$, $x \in \mathbb{R}^p$ for some $\lambda > 0$. Moreover, set $R(t) = CT(t)B$ and for $t_0 \geq 0$ and $x_0 \in X$ let $x(t; t_0, x_0)$ denote the mild solution of*

$$(5.1) \quad \begin{aligned} \dot{x}(t) &= Ax + Bf(t, Cx(t)), & t \geq t_0, \\ x(t_0) &= x_0. \end{aligned}$$

If $\|R\|, \lambda < 1$, then (5.1) has a unique, globally defined mild solution and there exist positive constants M and ε such that $\|x(t; t_0, x_0)\| \leq M e^{-\varepsilon(t-t_0)} \|x_0\|$ for all $x_0 \in X$ and $t \geq t_0 \geq 0$.

Remark 5.2. The assumptions made in Lemma 5.1 ensure that (5.1) admits a unique mild solution that can be continued to the right as long as it remains bounded (see Pazy [32]).

The proof of Lemma 5.1 is similar to the proof of Theorem 3.2 in Logemann [21] and is therefore omitted.

Now let us turn our attention to dynamical nonlinearities.

LEMMA 5.3. *Let X , $T(t)$, A , B , C , and $R(t)$ be as in Lemma 5.1. Suppose that $F: (LL^q(\mathbb{R}_+))^p \rightarrow (LL^q(\mathbb{R}_+))^p$ ($q = 1, 2, 3, \dots$; $q \neq \infty$) is causal, unbiased, and locally Lipschitz continuous. Then the following statements hold. (i) The equation*

$$(5.2) \quad x(t) = T(t)x_0 + \int_0^t T(t-\tau)BF(Cx(\cdot))(\tau) d\tau$$

admits for all $x_0 \in X$ a unique globally defined continuous solution $x(\cdot, x_0): [0, \infty) \rightarrow X$, which will be called the mild solution of

$$(5.3) \quad \dot{x}(t) = Ax(t) + BF(Cx(\cdot))(t), \quad x(0) = x_0;$$

and (ii) suppose that F additionally satisfies the condition

$$\|F(u)\|_q \leq \lambda_1 \|u\|_q + \lambda_2$$

for all $u \in (L^q(\mathbb{R}_+))^p$, where λ_1 and λ_2 are nonnegative constants; then the origin will be globally attractive (i.e., $\lim_{t \rightarrow \infty} x(t; x_0) = 0$ for all $x_0 \in X$) if

$$(5.4) \quad \|R\|_1 \lambda_1 < 1.$$

Proof. (i) It is clear that the mapping $u(\cdot) \mapsto F(Cu(\cdot))$ is causal, unbiased, and locally Lipschitz continuous. Hence it follows from Corollary 4.1.2 in Willems [38] that (5.2) has a unique solution x in $LL^q(\mathbb{R}_+, X)$. Since the right-hand side of (5.2) is continuous in t , we see that $x(t)$ is continuous as well.

(ii) Consider the equation

$$(5.5) \quad \begin{aligned} y(t) &= CT(t)x_0 + C \int_0^t T(t-\tau)BF(y(\cdot))(\tau) d\tau \\ &= CT(t)x_0 + \int_0^t R(t-\tau)F(y(\cdot))(\tau) d\tau. \end{aligned}$$

If $x(t) := x(t; x_0)$ is the solution of (5.2), then it is clear that $Cx(t)$ is a solution of (5.5). Using (5.4) and the fact that $CT(\cdot)x_0 \in (L^q(\mathbb{R}_+))^p$, it follows from the small-gain theorem that $Cx(\cdot) \in (L^q(\mathbb{R}_+))^p$ and hence $z(\cdot) := BF(Cx(\cdot)) \in L^q(\mathbb{R}_+, X)$. It remains to show that $w(t) := \int_0^t T(t-\tau)z(\tau) d\tau$ tends to zero as $t \rightarrow \infty$. By the exponential stability of $T(t)$, there exist positive constants N and γ such that $\|T(t)\| \leq N e^{-\gamma t}$ for all $t \geq 0$.

Suppose for a moment that $q \neq 1$, and define q' by $1/q' + 1/q = 1$:

$$\begin{aligned} \|w(t)\| &\leq N \left\{ \int_0^{t/2} e^{-\gamma(t-\tau)} \|z(\tau)\| d\tau + \int_{t/2}^t e^{-\gamma(t-\tau)} \|z(\tau)\| d\tau \right\} \\ &= N \left\{ \int_{t/2}^t e^{-\gamma\tau} \|z(t-\tau)\| d\tau + \int_{t/2}^t e^{-\gamma(t-\tau)} \|z(\tau)\| d\tau \right\} \\ &\leq N \left\{ \left(\int_{t/2}^t e^{-q'\gamma\tau} d\tau \right)^{1/q'} \left(\int_{t/2}^t \|z(t-\tau)\|^q d\tau \right)^{1/q} \right. \\ &\quad \left. + \int_{t/2}^t e^{-q'\gamma(t-\tau)} d\tau \right)^{1/q'} \left(\int_{t/2}^t \|z(\tau)\|^q d\tau \right)^{1/q} \Big\}. \end{aligned}$$

We obtain

$$\|w(t)\| \leq N \left\{ \left(\int_{t/2}^{\infty} e^{-q'\gamma\tau} d\tau \right)^{1/q'} \|z\|_q + \|e^{-\gamma}\|_{q'} \left(\int_{t/2}^{\infty} \|z(\tau)\|^q d\tau \right)^{1/q} \right\}.$$

Now $\lim_{t \rightarrow \infty} \int_{t/2}^{\infty} e^{-q'\gamma\tau} d\tau = 0$ and $\lim_{t \rightarrow \infty} \int_{t/2}^{\infty} \|z(\tau)\|^q d\tau = 0$ and thus $\lim_{t \rightarrow \infty} \|w(t)\| = 0$. A similar argument holds for the case where $q = 1$. \square

Remark 5.4. If in Lemma 5.3 (ii) $\lambda_2 = 0$, then it is not difficult to see that the origin of (5.2) is globally asymptotically stable; i.e., $\lim_{t \rightarrow \infty} x(t; x_0) = 0$ for all $x_0 \in X$, and for all $\varepsilon > 0$ there exists $\delta > 0$ such that $\|x_0\| \leq \delta$ implies $\|x(t; x_0)\| \leq \varepsilon$ for all $t \geq 0$.

In the following, let the plant be given by

$$(5.6) \quad \dot{x}(t) = Ax(t) + Bu(t), \quad x(t_0) = x_0, \quad y(t) = Cx(t),$$

where the linear operator A generates a strongly continuous semigroup $T(t)$ on a Banach space X , $B: \mathbb{R}^p \rightarrow X$, $C: X \rightarrow \mathbb{R}^p$ are bounded linear operators, (A, B) is exponentially stabilizable, and (C, A) is exponentially detectable. A minimal realization of the controller K_k defined in (3.3) is given by

$$(5.7) \quad \dot{z}(t) = \text{diag}_{1 \leq j \leq p} (k_j c_j) v(t), \quad z(t_0) = z_0, \quad w(t) = \Gamma z(t) + \Gamma \text{diag}_{1 \leq j \leq p} (k_j + c_j) v(t).$$

Let Φ be an operator mapping $(LL^q(\mathbb{R}_+))^p$ into itself. We will interpret Φ as a measurement nonlinearity in the feedback interconnection of (5.6) and (5.7) as follows:

$$(5.8) \quad u = w, \quad v = -\Phi(y).$$

THEOREM 5.5. *Suppose that $G(s) := C(sI - A)^{-1}B$ satisfies condition (3.1) (cf. Theorem 4.5), and define $x_c(t) := (x(t), z(t))'$. The following statements hold. (i) If Φ in (5.8) is given by N_φ , where φ satisfies (N1)–(N3), $\lambda_1 < 1$, $\lambda_2 = 0$, and $\|\Gamma^{-1}(\Gamma - D)\| < \frac{1}{2}(1 - \lambda_1)$ then for all sufficiently large gains k_j , $j = 1, \dots, p$, there exist positive constants M and ε (dependent on k) such that $\|x_c(t)\| \leq M e^{-\varepsilon(t-t_0)} \|x_c(t_0)\|$ for all $x_c(t_0) \in X \times \mathbb{R}^p$, $t \geq t_0 \geq 0$, i.e., the nonlinear feedback system given by (5.6)–(5.8) is globally exponentially stable; and (ii) if Φ in (5.8) satisfies (N4)–(N7), $q < \infty$, $\lambda_1 < 1$, and $\|\Gamma^{-1}(\Gamma - D)\| < \frac{1}{2}(1 - \lambda_1)$, then for all sufficiently large k_j , $j = 1, \dots, p$, the feedback system (5.6)–(5.8) is internally stable in the sense that the origin is globally attractive; i.e., $\lim_{t \rightarrow \infty} x_c(t) = 0$ for all $x_c(0) \in X \times \mathbb{R}^p$.*

Proof. (i) Set $B_k := \text{diag}_{1 \leq j \leq p} (k_j c_j)$ and $D_k := \Gamma \text{diag}_{1 \leq j \leq p} (k_j + c_j)$. A routine calculation then shows that

$$(5.9) \quad \dot{x}_c(t) = A_c x_c(t) - B_c \varphi_1(t, C_c x_c(t)),$$

where

$$A_c = A_c(k) := \begin{pmatrix} A - BD_k C & B\Gamma \\ -B_k C & 0 \end{pmatrix}, \quad B_c = B_c(k) := \begin{pmatrix} BD_k \\ B_k \end{pmatrix}, \quad C_c := (C \ 0).$$

We know from Theorem 3.5(i) that $\mathcal{F}[G, K_k]$ is H^∞ -stable for all sufficiently large k_j , $j = 1, \dots, p$, and thus, by a result of Jacobson and Nett [14] (cf. also Curtain [5]), $A_c(k)$ generates an exponentially stable semigroup $T_{c,k}(t)$ on $X \times \mathbb{R}^p$ for all large enough k_j , $j = 1, \dots, p$. It follows from the condition $\|\Gamma^{-1}(\Gamma - D)\| < \frac{1}{2}(1 - \lambda_1)$ and Lemma 3.8(ii) that for all sufficiently large k_j ($j = 1, \dots, p$) $\|\check{L}_k\|_{\lambda_1} < 1$. (Here we have used that $\lim_{k \rightarrow \infty} \|\check{L}_k^*\|_1 = 1$; cf. Logemann and Owens [22].) Finally, realize that $\check{L}_k(t) = C_c T_{c,k}(t) B_c(k)$, and apply Lemma 5.1 to (5.9).

(ii) Using the same arguments as in (i) and applying Lemma 5.3 instead of Lemma 5.1, we can prove the second claim. \square

Remark 5.6. It is well known that the retarded system of Example 3.3 admits an abstract state-space realization of the form (5.6), where the state space X is given by $M^2(-h, 0; \mathbb{R}^n) := \mathbb{R}^n \times L^2(-h, 0; \mathbb{R}^n)$. Using the notation of Example 3.3, let us assume that

$$(5.10) \quad \det \begin{pmatrix} sI - \hat{A}(s) & -B \\ C & 0 \end{pmatrix} \neq 0 \quad \text{for all } s \in \bar{\mathbb{C}}_0.$$

In particular, it follows from (5.10) that

$$\text{rk}(sI - \hat{A}(s) \ B) = n \quad \text{for all } s \in \bar{\mathbb{C}}_0$$

and

$$\text{rk} \begin{pmatrix} sI - \hat{A}(s) \\ C \end{pmatrix} = n \quad \text{for all } s \in \bar{\mathbb{C}}_0.$$

Hence the abstract state-space realization of the retarded system is exponentially stabilizable and exponentially detectable (see, e.g., Salamon [36]), and using Example 3.3 we see that under the extra assumption $\det(CB) \neq 0$, Theorem 5.5 applies to retarded systems.

Appendix 1.

Proof of Proposition 3.1. “Only if” Since G^{-1} admits a decomposition of form (3.1) with $H \in (H_\gamma^\infty)^{p \times p}$ for some $\gamma < 0$ we obtain

$$\begin{aligned} s(sG(s) - D^{-1}) &= s \left(\left(D + \frac{1}{s} H(s) \right)^{-1} - D^{-1} \right) \\ &= \left(D + \frac{1}{s} H(s) \right)^{-1} s \left(I - \left(D + \frac{1}{s} H(s) \right) D^{-1} \right) \\ &= - \left(D + \frac{1}{s} H(s) \right)^{-1} H(s) D^{-1}, \end{aligned}$$

which shows that $sG(s) - D^{-1} = O(1/s)$ as $|s| \rightarrow \infty$ in $\bar{\mathbb{C}}_\beta$, where $\beta \in (\gamma, 0)$ is arbitrary. To show that G has no zeros in $\bar{\mathbb{C}}_\beta$, pick holomorphic matrices $N, D \in \mathcal{H}(\mathbb{C}_\gamma)^{p \times p}$ such that N and D are right coprime and $G = ND^{-1}$. Then, trivially, $G^{-1} = DN^{-1}$ and by the right coprimeness of D and N it follows from the analyticity of G^{-1} in \mathbb{C}_γ that $\det(N)$ has no zeros in $\mathbb{C}_\gamma \cup \bar{\mathbb{C}}_\beta$.

“If” Setting $F(s) := (s + \gamma)G(s)$, $\gamma > |\beta|$, it follows from the assumption that

$$(A.1) \quad F(s) - D^{-1} = O(s^{-1}) \quad \text{as } |s| \rightarrow \infty \text{ in } \mathbb{C}_\beta$$

and, in particular,

$$(A.2) \quad \lim_{\substack{|s| \rightarrow \infty \\ s \in \mathbb{C}_\beta}} F(s) = D^{-1}.$$

Hence there exists $\rho > 0$ such that $F^{-1}(s)$ is bounded on $|s| > \rho$, $s \in \mathbb{C}_\beta$. Moreover, $F^{-1}(s) = (1/(s + \gamma))G^{-1}(s)$ and since G has no zeros in $\bar{\mathbb{C}}_\beta$, it follows that $F^{-1}(s)$ is bounded on $|s| \leq \rho$, $s \in \mathbb{C}_\beta$. Therefore F^{-1} is a bounded holomorphic function on \mathbb{C}_β , i.e., $F^{-1} \in (H_\beta^\infty)^{p \times p}$. Now realize that

$$(A.3) \quad \tilde{H}(s) := (s + \gamma)(F^{-1}(s) - D)$$

$$(A.4) \quad = (s + \gamma)F^{-1}(s)(D^{-1} - F(s))D.$$

It follows from (A.3) that \tilde{H} is holomorphic on \mathbb{C}_β and, furthermore, we obtain from (A.1), (A.2), and (A.4), using the boundedness of F^{-1} on \mathbb{C}_β , that \tilde{H} is bounded on \mathbb{C}_β ; hence $\tilde{H} \in (H_\beta^\infty)^{p \times p}$. Finally, we obtain

$$G^{-1}(s) = (s + \gamma)F^{-1}(s) = (s + \gamma)D + \tilde{H}(s) = sD + H(s),$$

where $H(s) := \gamma D + \tilde{H}(s)$. \square

Proof of Corollary 3.2. Since $A(s)$ is bounded on \mathbb{C}_γ for some $\gamma < 0$ there exists $\rho > 0$ such that

$$G(s) = \frac{1}{s} C \left(\sum_{j=0}^{\infty} \left(\frac{1}{s} A(s) \right)^j \right) B = \frac{1}{s} CB + \frac{1}{s} C \left(\sum_{j=1}^{\infty} \left(\frac{1}{s} A(s) \right)^j \right) B$$

for all $s \in \mathbb{C}_\gamma$ such that $|s| \geq \rho$. Hence

$$(A.5) \quad sG(s) - CB = O\left(\frac{1}{s}\right) \quad \text{as } |s| \rightarrow \infty \text{ in } \mathbb{C}_\gamma.$$

Moreover, since $\det(CB) \neq 0$, there exists an invertible matrix $Q \in \mathbb{C}^{n \times n}$ such that

$$Q^{-1}B = \begin{pmatrix} CB \\ 0 \end{pmatrix}, \quad CQ = (I_p \ 0).$$

Partition the matrix $Q^{-1}A(\cdot)Q$ as follows:

$$Q^{-1}A(\cdot)Q = \begin{pmatrix} A_{11}(\cdot) & A_{12}(\cdot) \\ A_{21}(\cdot) & A_{22}(\cdot) \end{pmatrix},$$

where $A_{11}(\cdot)$, $A_{12}(\cdot)$, $A_{21}(\cdot)$, and $A_{22}(\cdot)$ are matrices with entries in H_-^∞ of size $p \times p$, $p \times (n-p)$, $(n-p) \times p$, $(n-p) \times (n-p)$, respectively. As in Logemann [20], it follows that

$$(A.6) \quad \chi(s) = (-1)^p \det(CB) \det(sI - A_{22}(s)).$$

Now A_{22} is holomorphic and bounded on \mathbb{C}_α for some $\alpha < 0$, and therefore $\det(sI - A_{22}(s))$ has at most finitely many zeros in $\bar{\mathbb{C}}_\mu$ for any $\mu > \alpha$. Since, by assumption $\chi(s) \neq 0$ for all $s \in \bar{\mathbb{C}}_0$, we obtain, using (A.6),

$$(A.7) \quad \chi(s) \neq 0 \quad \text{for all } s \in \bar{\mathbb{C}}_\beta$$

for a negative β of sufficiently small modulus (without loss of generality, we may assume that $\gamma < \beta$). Let $G = ND^{-1}$ be a right coprime factorization over $\mathcal{H}(\mathbb{C}_\gamma)$ and use a well-known formula for the determinant of a four-block matrix to obtain

$$(A.8) \quad \begin{aligned} \chi(s) &= \det(sI - A(s)) \det(G(s)) \\ &= \frac{\det(sI - A(s))}{\det(D(s))} \det(N(s)). \end{aligned}$$

It is known that $\det(D(s))$ divides $\det(sI - A(s))$ (in $\mathcal{H}(\mathbb{C}_\gamma)$) (cf. Logemann [18]), and hence, by (A.8), we have that $\det(N)$ divides χ (in $\mathcal{H}(\mathbb{C}_\gamma)$). Thus, by (A.7),

$$(A.9) \quad \det(N(s)) \neq 0 \quad \text{for all } s \in \bar{\mathbb{C}}_\beta.$$

The claim now follows from (A.5), (A.9), and Proposition 3.1. \square

Appendix 2.

Proof of Lemma 3.8. An elementary computation shows that

$$(A.10) \quad L_k - L_k^* = \{[I - J_k(I - \Gamma^{-1}D) + P_k]^{-1} - I\}L_k^*,$$

where

$$J_k(s) := \text{diag}_{1 \leq j \leq p} \left(\frac{s}{s + k_j} \right)$$

and

$$P_k(s) := \text{diag}_{1 \leq j \leq p} \left(\frac{s}{(s + k_j)(s + c_j)} \right) \left\{ \Gamma^{-1} H(s) + \text{diag}_{1 \leq j \leq p} (c_j) (I - \Gamma^{-1} D) \right\}.$$

Note that we can factorize P_k as $P_k = J_k Q$, where

$$Q(s) := \text{diag}_{j=1, \dots, p} \left(\frac{1}{s + c_j} \right) \{ \Gamma^{-1} H(s) + \text{diag} (c_j) (I - \Gamma^{-1} D) \}.$$

Using a result by Mossaheb [27] (cf. also Logemann [18]), we see that $\check{Q} e^{\varepsilon \cdot} \in (L^1(\mathbb{R}_+))^{p \times p}$ for all sufficiently small $\varepsilon > 0$. Moreover, we have

$$(A.11) \quad \lim_{k \rightarrow \infty} \|\check{P}_k\|_1 = \lim_{k \rightarrow \infty} \|\check{J}_k * \check{Q}\|_1 = 0,$$

which can be derived using the equation $\check{J}_k = \delta_0 I - \text{diag}_{1 \leq j \leq p} (k_j e^{-k_j \cdot})$ and Lemma A.1. In the case where $\Gamma = D$, we obtain from (A.10), by taking inverse Laplace transforms,

$$(A.12) \quad \check{L}_k - \check{L}_k^* = \{(\delta_0 I + \check{P}_k)^{-1} - \delta_0 I\} * \check{L}_k^*.$$

It follows from (A.11) that the inverse of $\delta_0 I + \check{P}_k$ exists (in the Banach algebra $\mathcal{A}^{p \times p}$) if $\min_{1 \leq j \leq p} (k_j)$ is sufficiently large. Hence (A.12) makes sense for large $k_j, j = 1, \dots, p$. Part (i) of the lemma now follows from (A.12), (A.11), and the fact that

$$(A.13) \quad \lim_{k \rightarrow \infty} \|\check{L}_k^*\|_1 = 1$$

(cf. Logemann and Owens [22] for (A.13)).

To prove part (ii), set $M_k := J_k(I - \Gamma^{-1} D)$ and realize that $\|\check{M}_k\|_{\mathcal{A}} \leq \|\check{J}_k\|_{\mathcal{A}} \|I - \Gamma^{-1} D\| \leq 2\varepsilon < 1$. Taking inverse Laplace transforms, using (A.11), and employing the fact that $\mathcal{A}^{p \times p}$ is a Banach algebra, it follows from (A.10) that

$$\check{L}_k - \check{L}_k^* = \{[\delta_0 I - (\check{M}_k - \check{P}_k)]^{-1} - \delta_0 I\} * \check{L}_k^* = \left(\sum_{n=1}^{\infty} (\check{M}_k - \check{P}_k)^n \right) * \check{L}_k^*$$

for all sufficiently large $k_j, j = 1, \dots, p$. Moreover,

$$\|\check{L}_k - \check{L}_k^*\|_1 \leq \left(\sum_{n=1}^{\infty} (2\varepsilon + \|\check{P}_k\|_1)^n \right) \|\check{L}_k^*\|_1 = \frac{2\varepsilon + \|\check{P}_k\|_1}{1 - (2\varepsilon + \|\check{P}_k\|_1)} \|\check{L}_k^*\|_1$$

for all sufficiently large $k_j, j = 1, \dots, p$. We obtain, by using (A.13) and (A.11),

$$\limsup_{k \rightarrow \infty} \|\check{L}_k - \check{L}_k^*\|_1 \leq \frac{2\varepsilon}{1 - 2\varepsilon},$$

which is (ii). \square

LEMMA A.1. Set $e_k(t) := k e^{-kt} \theta(t)$, $t \geq 0$, $k \geq 0$. Then $\lim_{k \rightarrow \infty} \|e_k * f - f\|_1 = 0$ for all $f \in L^1(\mathbb{R}_+)$.

Remark A.2. Note that e_k is not a so-called approximate identity or Dirac sequence, because the support of e_k does not shrink to $\{0\}$ as $k \rightarrow \infty$.

Proof of Lemma A.1. In the following set $f(t) := 0$ for all $t < 0$

$$\begin{aligned}
 \|e_k * f - f\|_1 &= \int_0^\infty \left| \int_0^t e_k(t-s)f(s) ds - f(t) \right| dt \\
 (A.14) \qquad &= \int_0^\infty \left| \int_0^\infty e_k(\tau)(f(t-\tau) - f(t)) d\tau \right| dt \\
 &\leq \int_0^\infty e_k(\tau) \left\{ \int_0^\infty |f(t-\tau) - f(t)| dt \right\} d\tau.
 \end{aligned}$$

It is well known that for a given $\varepsilon > 0$ there exists $\delta > 0$ such that $\int_0^\infty |f(t-\tau) - f(t)| dt \leq \varepsilon$ for all $\tau \in [0, \delta]$. Hence it follows from (A.14) that

$$\begin{aligned}
 \|e_k * f - f\|_1 &\leq \int_0^\delta e_k(\tau) \left\{ \int_0^\infty |f(t-\tau) - f(t)| dt \right\} d\tau + 2\|f\|_1 \int_\delta^\infty e_k(\tau) d\tau \\
 &\leq \varepsilon + 2\|f\|_1 e^{-k\delta} \leq 2\varepsilon
 \end{aligned}$$

for all sufficiently large k .

Appendix 3. In this appendix we will present the most important results on stabilizability of infinite-dimensional systems. We use the same notation as in § 4.

DEFINITION A.3. System (A, B) is α -exponentially stabilizable if there exists a bounded linear operator $F \in L(X, \mathbb{R}^p)$ such that the semigroup $T_{BF}(t)$ generated by $A + BF$ satisfies $\|T_{BF}(t)\| \leq M e^{\beta t}$ for some $M \geq 1$ and $\beta < \alpha$. System (A, B) is exponentially stabilizable if it is 0-exponentially stabilizable. System (C, A) is α -exponentially detectable if there exists a bounded linear operator $K \in L(\mathbb{R}^p, X)$ such that the semigroup $T_{KC}(t)$ generated by $A + KC$ satisfies $\|T_{KC}(t)\| \leq M e^{\beta t}$ for some $M \geq 1$ and $\beta < \alpha$. System (C, A) is exponentially detectable if it is 0-exponentially detectable.

LEMMA A.4. Suppose that the underlying space X is reflexive. Then system (A, B) is α -exponentially stabilizable if and only if (B^*, A^*) is α -exponentially detectable.

We now have the following important theorem.

THEOREM A.5. The following conditions are equivalent. (i) System (A, B) is α -exponentially stabilizable; and (ii) the state space can be decomposed in two semigroup-invariant subspaces $X = X_s \oplus X_u$, where X_s and X_u satisfy

$$\|T(t)|_{X_s}\| \leq M e^{\beta t}; \quad M \geq 1, \quad \beta < \alpha,$$

$$\dim(X_u) < \infty,$$

$$\sigma(A|_{X_u}) = \sigma(A) \cap \bar{C}_\alpha = \sigma_p(A) \cap \bar{C}_\alpha,$$

—The finite-dimensional system $(A|_{X_u}, P_{X_u}B)$ is controllable, where P_{X_u} is the projection on X_u along X_s .

For the proof, see Desch and Schappacher [6], Nefedov and Sholokhovitch [29], Jacobson and Nett [14], or Curtain [5].

It is easy to show that X_u is the span of all unstable (generalized) eigenvectors of A .

The following theorem is used frequently in § 4.

THEOREM A.6. Assume that (A, B) is α -exponentially stabilizable and let $Q \in L(X)$ be compact. Then $A + Q$ generates an α -exponentially stable semigroup if and only if $\sigma_p(A + Q) \cap \bar{C}_\alpha = \emptyset$.

Proof. It follows from Theorem A.5 that the essential exponential growth bound $\omega_e(T(\cdot))$ of $T(t)$ (cf., e.g., Nagel [28]) satisfies $\omega_e(T(\cdot)) < \alpha$. Since Q is compact, we have that $\omega_e(T_Q(\cdot)) = \omega_e(T(\cdot)) < \alpha$. Let $\omega(T_Q(\cdot))$ denote the exponential growth

bound of the semigroup $T_Q(t)$. We must prove that $\omega(T_Q(\cdot)) < \alpha$. Let us assume the contrary. Then $\omega(T_Q(\cdot)) > \omega_e(T_Q(\cdot))$ and we can show, as in [28, p. 74], that there exists $\lambda \in \sigma_p(A+Q)$ satisfying $\operatorname{Re}(\lambda) = \omega(T_Q(\cdot)) \geq \alpha$. This leads to a contradiction because $\sigma_p(A+Q) \cap \bar{\mathbb{C}}_\alpha = \emptyset$ by assumption.

REFERENCES

- [1] S. P. BANKS AND F. ABBASI-GHELMANSARAI, *Realization theory and the infinite-dimensional root locus*, Internat. J. Control, 38 (1983), pp. 589–606.
- [2] S. BOYD AND C. A. DESOER, *Subharmonic functions and performance bounds on linear time-invariant feedback systems*, IMA J. Math. Control Inform., 2 (1985), pp. 153–170.
- [3] C. I. BYRNES AND D. S. GILLIAM, *Asymptotic properties of root locus for distributed parameter systems*, in Proc. 27th Conference on Decision and Control, Austin, TX, 1988, pp. 45–51.
- [4] F. M. CALLIER AND J. WINKIN, *Distributed system transfer functions of exponential order*, Internat. J. Control, 43 (1986), pp. 1353–1373.
- [5] R. F. CURTAIN, *Equivalence of input-output stability and exponential stability for infinite-dimensional systems*, Math. Systems Theory, 21 (1988), pp. 19–48.
- [6] W. DESCH AND W. SCHAPPACHER, *Spectral properties of finite-dimensional perturbed linear semigroups*, J. Differential Equations, 59 (1985), pp. 80–102.
- [7] C. A. DESOER AND Y.-T. WANG, *On the generalized Nyquist criterion*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 187–196.
- [8] G. F. FRANKLIN, J. D. POWELL, AND A. EMAMI-NAEINI, *Feedback Control of Dynamic Systems*, Addison-Wesley, Reading, MA, 1986.
- [9] J. S. FREUDENBERG AND D. P. LOOZE, *A sensitivity tradeoff for plants with time delay*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 99–104.
- [10] P. R. HALMOS, *A Hilbert Space Problem Book*, 2nd ed., Springer-Verlag, New York, 1982.
- [11] C. J. HARRIS AND J. M. E. VALENCIA, *The Stability of Input-Output Dynamical Systems*, Academic Press, London, 1983.
- [12] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, Colloq. Publ. Amer. Math. Soc., 1957.
- [13] C. A. JACOBSON, *The structure of exponential stabilization of a class of linear distributed parameter systems*, Ph.D. thesis, Rensselaer Polytechnic Institute, Troy, New York, 1986.
- [14] C. A. JACOBSON AND C. N. NETT, *Linear state-space systems in infinite-dimensional space: The role and characterization of joint stabilizability/detectability*, IEEE Trans. Automat. Control, AC-33 (1988), pp. 541–549.
- [15] T. T. JUSSILA AND H. N. KOIVO, *Tuning of multivariable PI-controllers for unknown delay-differential systems*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 364–368.
- [16] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1966.
- [17] T. KOBAYASHI, *A digital PI-controller for distributed parameter systems*, SIAM J. Control Optim., 26 (1988), pp. 1399–1414.
- [18] H. LOGEMANN, *Funktionentheoretische Methoden in der Regelungstheorie unendlich-dimensionaler Systeme*, Ph.D. thesis, Institut für Dynamische Systeme, Universität Bremen, Bremen, Germany, 1986.
- [19] ———, *On the Nyquist criterion and robust stabilization for infinite-dimensional systems*, in Robust Control of Linear Systems and Nonlinear Control, M. A. Kaashoek, J. H. van Schuppen, and A. C. M. Ran, eds., Birkhäuser, Boston, 1990, pp. 627–633.
- [20] ———, *Adaptive exponential stabilization for a class of nonlinear retarded processes*, Math. Control Signals Systems, 3 (1990), pp. 255–269.
- [21] ———, *Circle criteria, small gain conditions and internal stability for infinite-dimensional systems*, Automatica, 27 (1991), pp. 677–690.
- [22] H. LOGEMANN AND D. H. OWENS, *Robust high-gain feedback control of infinite-dimensional minimum-phase systems*, IMA J. Math. Control Inform., 4 (1987), pp. 195–220.
- [23] ———, *Robust and adaptive high-gain control of infinite-dimensional systems*, in Analysis and Control of Nonlinear Systems, C. I. Byrnes, C. F. Martin, and R. E. Saeks, eds., North-Holland, Amsterdam, 1988, pp. 35–44.
- [24] ———, *Low-gain control of unknown infinite-dimensional systems: A frequency-domain approach*, Dynamics Stability Systems, 4 (1989), pp. 13–29.
- [25] A. I. MEES, *Dynamics of Feedback Systems*, John Wiley, Chichester, UK, 1981.
- [26] R. K. MILLER, *Nonlinear Volterra Integral Equations*, Benjamin Cummings, Menlo Park, CA, 1971.

- [27] S. MOSSAHEB, *On the existence of right-coprime factorizations for functions meromorphic in a half-plane*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 550–551.
- [28] R. NAGEL, ED., *One Parameter Semigroups of Positive Operators*, Lecture Notes in Math. 1184, Springer-Verlag, Berlin, 1986.
- [29] S. A. NEFEDOV AND F. A. SHOLOKHOVICH, *A criterion for the stabilizability of dynamical systems with finite-dimensional input*, Differentsial'nye Uravneniya, 22 (1986), pp. 163–166.
- [30] C. N. NETT, C. A. JACOBSON, AND M. J. BALAS, *A connection between state space and doubly coprime fractional representations*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 831–832.
- [31] D. H. OWENS AND A. CHOTAI, *High performance controllers for unknown multivariable systems*, Automatica, 18 (1982), pp. 583–587.
- [32] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
- [33] S. POHJOLAINEN, *Computation of transmission zeros for distributed parameter systems*, Internat. J. Control, 33 (1981), pp. 199–212.
- [34] ———, *Robust multivariable PI-controller for infinite-dimensional systems*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 17–30.
- [35] W. RUDIN, *Real and Complex Analysis*, 2nd ed., McGraw-Hill, New York, 1984.
- [36] D. SALAMON, *Control and Observation of Neutral Systems*, Research Notes in Math. 91, Pitman, Boston, 1984.
- [37] M. VIDYASAGAR, H. SCHNEIDER, AND B. A. FRANCIS, *Algebraic and topological aspects of feedback stabilization*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 880–894.
- [38] J. C. WILLEMS, *The Analysis of Feedback Systems*, MIT Press, Cambridge, MA, 1971.
- [39] J. ZABCZYK, *On decomposition of generators*, SIAM J. Control Optim., 16 (1978), pp. 523–534.
- [40] H. J. ZWART, *Geometric Theory for Infinite-Dimensional Systems*, Lecture Notes in Control and Inform. Sci., Vol. 115, Springer-Verlag, Berlin, 1989.

STOCHASTIC DISCRETE OPTIMIZATION*

DI YAN† AND H. MUKAI†

Abstract. In this paper a stochastic search method is proposed for finding a global solution to the stochastic discrete optimization problem in which the objective function must be estimated by Monte Carlo simulation. Although there are many practical problems of this type in the fields of manufacturing engineering, operations research, and management science, there have not been any nonheuristic methods proposed for such discrete problems with stochastic infrastructure. The proposed method is very simple, yet it finds a global optimum solution. The method exploits the randomness of Monte Carlo simulation and generates a sequence of solution estimates. This generated sequence turns out to be a nonstationary Markov chain, and it is shown under mild conditions that the Markov chain is strongly ergodic and that the probability that the current solution estimate is global optimum converges to one. Furthermore, the speed of convergence is also analyzed.

Key words. stochastic optimization, discrete parameters, Monte Carlo simulation, global optimization, Markov chain

AMS(MOS) subject classifications. 62L99, 90B22, 90C99

1. Introduction. In the fields of manufacturing engineering, operations research, and management science, we often find a discrete optimization problem in which an objective function g is minimized over a nonempty discrete finite feasible set S :

$$(1.1) \quad \min \{g(s) | s \in S\},$$

where $g: S \rightarrow \mathbb{R}$ and $S = \{s_1, s_2, \dots, s_k\}$. In such a problem, the discreteness of the feasible set is the most salient characteristic, while the finiteness of the feasible set is mere convenience in theory without imposing any serious limitations in practice since the size can be made larger as needed.

In examining many practical problems of the form (1.1), we note that the objective function $g(s)$ is often the expectation of the performance of a system that is subject to stochastic phenomena. Hence we give $g(s)$ the following stochastic infrastructure:

$$(1.2) \quad g(s) = E[h(s, Y(s))],$$

where E denotes the expectation, h is a function of s and y , and $Y(s)$ is a random vector dependent on s . Hence $h(s, y)$ may represent the performance of the system when the parameter is set to s and the sample of random vector $Y(s)$ is y . In such problems, a closed-form formula is often not available for the objective function $g(s)$, and one is forced to estimate $g(s)$ by Monte Carlo-type simulation. For this purpose, we write a computer program for simulating the system and drive it by random number generators. Here the simulation program represents $h(s, y)$ and the random number generators represent $Y(s)$.

As an example, consider a queueing network consisting of stations with buffers. Such a network may represent an assembly line in the manufacturing industry, a network of processors for a parallel computer, or a communication network for message packets. Suppose that the optimal sizes of buffers are sought. Hence vector s represents buffer sizes at different stations. Random vector Y may represent the lengths of service time at different stations and the lengths of transfer time from station to station. The function $h(s, y)$ is often set to the amortized cost of buffers minus the revenues due

* Received by the editors June 5, 1989; accepted for publication (in revised form) April 4, 1991. This research was supported in part by United States Air Force grant AFOSR-89-0518.

† Department of Systems Science and Mathematics, Washington University, St. Louis, Missouri 63130.

to the processing speed. Therefore one tries to minimize $g(s) = E[h(s, Y)]$. In this case it is relatively easy to write a simulation program for $h(s, y)$ and to drive y by random number generators representing Y , but it is very time consuming to write an explicit formula for $h(s, y)$ in closed form. Furthermore when a multistation multiserver queueing network is subject to some recycling and a complex priority system, it would be impossible [3, p. 455] to model the network analytically and to obtain an explicit formula for $h(s, y)$.

As another example, consider a telephone installation problem in which several types of telephones are installed at an airport. One type of telephone may accept coins only, another may accept credit cards only, and yet another may accept both. Here vector s represents the numbers of telephones in different categories, and random vector $Y(s)$ may represent the length of time period between customer arrivals, and the length of time a customer spends at a telephone. We note that the latter depends on the type of telephone since credit card calls tend to take extra time for dialing the card number and for verification by a computer. Here the performance $h(s, y)$ may represent the average waiting time for a customer and the amortized cost of telephones.

In the past, a paper [5] considered a specific problem of buffer size selection in a serial production line and proposed a very heuristic method without any convergence analysis.

In the absence of any computational methods with theoretical underpinnings that are specifically designed to solve the discrete optimization problem with stochastic infrastructure, one might naively combine Monte Carlo simulation and a discrete optimization technique developed for the case in which the objective function is readily deterministically computable. The most straightforward approach would be to replace the objective function $g(s)$ by its estimate $\hat{g}_\ell(s)$ based on ℓ simulation experiments. For example, $\hat{g}_\ell(s)$ may be defined as $\hat{g}_\ell(s) = (1/\ell) \sum_{i=1}^{\ell} h(s, y_i(s))$, where $y_1(s), \dots, y_\ell(s)$ denote ℓ samples of $Y(s)$. In this case, it is not obvious how large the sample size ℓ should be to guarantee the convergence of the optimization technique.

In this paper, we take a different approach and propose a new method for solving the discrete optimization problem with stochastic infrastructure. In the above naive approaches, stochastic elements are suppressed by taking large sample sizes ℓ . Here we treat stochastic elements as they are in our proposed method. Moreover, we exploit the stochastic property of simulation in finding global optimal solution. In the proposed method, the new solution candidate is compared with an absolute scale. Given a practical situation, one can often guess an approximate range (a, b) for the stochastic objective function $h(s, Y(s))$ from experience. When such experience is not available, we can always run simulation experiments for different values of s and obtain an approximate range (a, b) for the stochastic objective function $h(s, Y(s))$. Then we may define a random variable $\Theta(a, b)$, which is uniformly distributed over the interval (a, b) , and we use this as a scale against which $h(s, Y(s))$ is measured. Thus we convert the original minimization problem (1.1) into the following probability maximization problem:

$$\max \{ \text{Prob} [h(s, Y(s)) \leq \Theta(a, b)] \mid s \in S \}.$$

Hence our strategy is to solve this maximization problem in place of the original problem (1.1). Here we must note that the probability being maximized is not numerically computable. Rather the probability is implicitly represented in the probabilistic simulation, and we exploit that implicit representation in designing a method for solving this maximization problem. In § 2, we exactly state the discrete optimization problem with stochastic infrastructure. Then, in § 3, we translate the original minimization problem into the probability maximization problem and discuss the relationship

between the two problems. In § 4, we list assumptions and describe our proposed method, which depends on a sequence $\{M_k\}$ of parameters for generating a sequence $\{X_k\}$ of solution estimates. Then, in § 5, we freeze the parameter M_k to a constant M , analyze the resulting stationary Markov chain, and obtain its stationary probability distribution $\pi(M)$ as a function of M . In § 6, we study the behavior of the stationary probability distribution $\pi(M)$ as M goes to infinity. Then, in § 7, we study the proposed method when the sequence $\{M_k\}$ of parameters monotonically increases to infinity, analyze the resulting nonstationary Markov chain, establish that the chain is strongly ergodic, and show that the probability that the solution estimate X_k is in the global optimum set goes to one as the iteration k goes to infinity. Furthermore, in § 8, we investigate the rate of convergence of the proposed method. Then the paper ends with some conclusions in § 9. In the interest of readability, all the proofs are relegated to the Appendix.

2. Problem structure and assumptions. Consider problem (1.1). Instead of local solutions, we seek global solutions to this problem. We denote the (global) optimum set by

$$(2.1) \quad S^* = \{s \in S \mid g(s) \leq g(s'), \forall s' \in S\}.$$

The objective function is now assumed to have the structure of (1.2), where $h(s, y)$ is a measurable function of s and y , and $Y(s)$ is a random vector defined on probability space (Ω_s, F_s, P_s) for each $s \in S$. We observe that the random variable $H(s)$ defined by

$$(2.2) \quad H(s) = h(s, Y(s))$$

is well defined on (Ω_s, F_s, P_s) for each $s \in S$. We further note that the probability distribution for $Y(s)$ is often known a priori and simple (e.g., the exponential distribution and the Gaussian distribution), but that the probability distribution for $H(s)$ is not generally known a priori and complex.

We shall assume throughout the paper that the variance of $H(s)$ is finite for each $s \in S$.

ASSUMPTION A0. We have that

$$(2.3) \quad E[H(s)^2] < \infty, \forall s \in S.$$

We would like to point out that this finite variance property is the only property that we demand of the problem.

3. Translation to a maximization problem. Consider the problem defined in the preceding sections. Given an $s \in S$, we can only obtain samples $\{h_i(s)\}$ of random variable $H(s)$ via simulation. Hence, given two elements, s and s' , from S , we cannot decide which element is better than the other based on a finite number of samples $\{h_i(s)\}$ and $\{h_i(s')\}$. However, these samples do contain some information about their underlying stochastic structures, and we need to extract it. To accomplish this, we propose to measure $H(s)$ against a stochastic ruler—another random variable. With this idea, we will be able to translate the original minimization problem into a maximization problem of a certain probability. As a stochastic ruler, we have selected a uniformly distributed random variable in this paper, but other choices are possible.

Let $\Theta(a, b)$ denote the random variable uniformly distributed between a and b provided $a < b$. Here a and b , respectively, represent in a loose sense a lower and an upper bound for $\{H(s) \mid s \in S\}$. For example, if it is known that $a' \leq H(s) \leq b'$ for all $s \in S$, then a and b may be, respectively, set to a' and b' .

Now we compare $H(s)$ with $\Theta(a, b)$ and let

$$(3.1) \quad P(s, a, b) = P[H(s) \leq \Theta(a, b)].$$

Here the random variable $H(s)$ representing the objective value is compared with a stochastic ruler $\Theta(a, b)$. We can intuitively see that minimizing $g(s) = E[H(s)]$ is equivalent to maximizing the probability $P(s, a, b)$ provided the interval (a, b) is sufficiently wide. Hence we consider the following maximization problem:

$$(3.2) \quad \max \{P(s, a, b) \mid s \in S\}.$$

The global optimum solution set for this maximization problem is

$$(3.3) \quad S^*(a, b) = \{s \in S \mid P(s, a, b) \geq P(s', a, b) \forall s' \in S\}.$$

The following theorem rigorously delineates the relationship between the original minimization problem (1.1) and the above maximization problem (3.2).

THEOREM 3.1. *There exist real numbers \bar{a} and \bar{b} such that $\bar{a} < \bar{b}$ and for any $a < \bar{a}$ and any $b > \bar{b}$, the following conclusions hold:*

- (1) *If $g(s) < g(s')$ then $P(s, a, b) > P(s', a, b)$,*
- (2) *$0 < P(s, a, b) < 1$, for all $s \in S$,*
- (3) *$S^*(a, b) \subset S^*$ and $S^*(a, b) \neq \emptyset$.*

Theorem 3.1 states that maximization problem (3.2) has at least one solution and that any solution of that problem is a solution of the original minimization problem (1.1) provided the interval (a, b) is sufficiently large. The natural question now is: Does the converse hold? We partly answer that question in the next theorem.

THEOREM 3.2. *Suppose there exist reals $a(s)$ and $b(s)$ such that*

$$(3.4) \quad a(s) \leq H(s) \leq b(s) \quad \text{w.p.1.}$$

If $a < \min \{a(s) \mid s \in S\}$ and $b > \max \{b(s) \mid s \in S\}$, then $S^(a, b) = S^*$.*

In closing this section, we note that assumption (3.4) of Theorem 3.2 is not necessary for the method proposed in the next section, which solves the maximization problem (3.2). Indeed, conclusion (3) of Theorem 3.1 guarantees that a solution exists for the latter problem (3.2) and that its solution is also a solution for the original problem (1.1).

Furthermore we note that probability (3.1) is not explicitly available. Hence it is not possible to apply any conventional discrete optimization techniques to (3.2) in a straightforward manner. However, probability (3.1) is implicitly available in the simulation of $H(s)$ and $\Theta(a, b)$, and we exploit this fact to construct a method for solving (3.2) in the next section.

4. Computational method. In this section we present a stochastic algorithm for solving the discrete optimization problem discussed in §§ 1 and 2. The basic idea of our approach is to make use of probabilistic simulation in constructing a nonstationary Markov chain that converges to a global solution to the problem. Our approach is related to, but different from, the technique of simulated annealing.

First, while the objective value in simulated annealing is assumed to be exactly computable, our method accepts samples of the objective value generated by probabilistic simulation. Second, while the objective value at a new solution candidate is compared with that of the current solution candidate in simulated annealing, the objective value at a new solution candidate is compared against a probabilistic ruler in our method.

In the following, we list definitions and assumptions we will make use of in the rest of this paper. First, to structure our search for an optimal solution in the feasible set S , we introduce the concept of neighbors in S .

DEFINITION 4.1. For each $s \in S$, there exists a subset $N(s)$ of $S - \{s\}$, which is called *the set of neighbors* of s .

Our search is organized in such a way that the next solution candidate is found among the neighbors of the present candidate. Hence, to ensure that our search will eventually cover all the elements of S , we make the following assumption about the system N of neighbors.

ASSUMPTION A1. For any pair (s, s') in $S \times S$, s' is *reachable* from s ; i.e., there exists a finite sequence, $\{n_i\}_{i=0}^\ell$ for some ℓ , such that

$$s_{n_0} = s, \quad s_{n_\ell} = s', \quad s_{n_{i+1}} \in N(s_{n_i}), \quad i = 0, 1, 2, \dots, \ell - 1.$$

Now we impose a stochastic structure to the selection of a candidate among the neighbors by the following function R . Given an $s \in S$, a candidate is selected among $N(s)$ such that the probability of selecting a neighbor $s' \in N(s)$ is equal to $R(s, s')$.

DEFINITION 4.2. A function $R: S \times S \rightarrow [0, 1]$ is said to be a *transition probability* for S and N if

- (1) $R(s, s') > 0 \Leftrightarrow s' \in N(s)$ and
- (2) $\sum_{s' \in S} R(s, s') = 1$.

Statement (1) says that every neighbor and only neighbors alone are given positive probability to be selected as a candidate. Statement (2) says the probabilities of selecting all the neighbors must add up to one. The simplest way to define $R(s, s')$ is perhaps to distribute the probability uniformly over $N(s)$, i.e., $R(s, s') = 1/|N(s)|$ for $s' \in N(s)$, and $R(s, s') = 0$ for $s' \notin N(s)$, where $|N(s)|$ denotes the number of elements in $N(s)$. However, given a particular problem, it may be more advantageous to skew the distribution.

Now we introduce the following simplification.

ASSUMPTION A2. The neighbor system N and the transition probability R for S are *symmetric*, i.e.,

- (1) $s' \in N(s) \Leftrightarrow s \in N(s')$ and
- (2) $R(s, s') = R(s', s)$.

In the algorithm, we make use of a sequence of positive integers tending to infinity.

ASSUMPTION A3. A sequence $\{M_k\}$ of positive integers satisfies $M_k \rightarrow \infty$ as $k \rightarrow \infty$.

Aside from N , R , and $\{M_k\}$ defined above, the proposed stochastic algorithm requires parameters, a and b , and an initial guess $s_0 \in S$ for the optimal solution.

THE STOCHASTIC ALGORITHM.

Data: $N, R, \{M_k\}, a, b, s_0 \in S$.

Step 0: Set $X_0 = s_0$ and $k = 0$.

Step 1: Given $X_k = s$, choose a candidate Z_k from $N(s)$ with probability distribution

$$P[Z_k = s' / X_k = s] = R(s, s'), \quad s' \in N(s).$$

Step 2: Given $Z_k = s'$, set

$$X_{k+1} = \begin{cases} Z_k, & \text{with probability } p_k, \\ X_k, & \text{with probability } (1 - p_k), \end{cases}$$

where

$$p_k = \{P[H(s') \leq \Theta(a, b)]\}^{M_k} = \{P(s', a, b)\}^{M_k}.$$

Remark. Since we are interested in cases in which the probability $P(s', a, b)$ given above in Step 2 is not explicitly computable, we suggest a subalgorithm for implementing Step 2 immediately following the algorithm.

Step 3: Set $k = k + 1$ and go to Step 1.

The implementation of Step 2 of the above algorithm may be accomplished by the following subalgorithm where $P(s', a, b)$ need not be computed.

SUBALGORITHM FOR STEP 2.

Given $Z_k = s'$, draw a sample $h(s')$ from $H(s')$, or equivalently draw a sample $y(s')$ from $Y(s')$ using random number generators and then compute $h(s', y(s'))$ by running one simulation experiment. Next draw a sample θ from $\Theta(a, b)$. If $h(s') = h(s', y(s')) > \theta$ then set $X_{k+1} = X_k$; otherwise draw another sample $h(s')$ from $H(s')$ (or equivalently draw another sample $y(s')$ and compute $h(s', y(s'))$ by running another experiment), and compare this against another sample θ from $\Theta(a, b)$. If $h(s') = h(s', y(s')) > \theta$, then set $X_{k+1} = X_k$; otherwise continue to draw and compare. If all M_k tests, $h(s') = h(s', y(s')) > \theta$, fail, then we accept the candidate Z_k and set $X_{k+1} = Z_k = s'$.

The random process $\{X_k\}$ produced by the Stochastic Algorithm is a discrete-time Markov chain defined over states S , and its state transition probabilities are given by

$$(4.1) \quad \begin{aligned} P_{ss'}(M_k) &= P[X_{k+1} = s' / X_k = s] \\ &= \begin{cases} R(s, s') \{P(s', a, b)\}^{M_k}, & \text{if } s' \in N(s); \\ 1 - \sum_{s'' \in N(s)} R(s, s'') \{P(s'', a, b)\}^{M_k}, & \text{if } s' = s; \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

We make use of the state transition probability matrix, which is a matrix consisting of the above probabilities:

$$(4.2) \quad P(M_k) = (P_{ss'}(M_k)).$$

We also note that $P(M_k)$ is a stochastic matrix [6].

5. Analysis for the stationary process. In this section, we suspend Assumption A3 and set M_k to a positive constant integer M to investigate the stationary behavior of the algorithm in § 4. Indeed, the Markov chain then becomes stationary since the state transition probability (4.1) becomes independent of k .

For each $s \in S$, define

$$(5.1) \quad \pi_s(M) = \frac{\{P[H(s) \leq \Theta(a, b)]\}^M}{\sum_{s' \in S} \{P[H(s') \leq \Theta(a, b)]\}^M} = \frac{\{P(s, a, b)\}^M}{\sum_{s' \in S} \{P(s', a, b)\}^M}.$$

We will show in the next theorem that $\{\pi_s(M) | s \in S\}$ is the stationary probability distribution for the stationary Markov chain X_k generated by the algorithm with M_k set to M .

We now assume for the rest of the paper that parameters a and b are selected so as to satisfy the conditions in Theorem 3.1.

THEOREM 5.1. *The vector $\pi(M)$ consisting of $\pi_s(M)$ in (5.1) represents the stationary probability distribution for the Markov chain $\{X_k\}$ defined by (4.1) with $M_k = M$. Furthermore, this vector is the left eigenvector with eigenvalue one for the state transition probability matrix $P(M)$ defined in (4.2), i.e.,*

$$(5.2) \quad \pi(M)P(M) = \pi(M).$$

6. The limiting behavior of the stationary distribution. In this section, we investigate the behavior of the stationary probability distribution $\{\pi_s(M) | s \in S\}$ found in the preceding section as M goes to infinity.

DEFINITION 6.1. Given a finite set S , the set $\Pi(S)$ of positive unit vectors is called the set of probability vectors for S , below:

$$\Pi(S) = \{\pi \in [0, 1]^\kappa \mid \pi_s \geq 0, \|\pi\| = \sum_{s \in S} \pi_s = 1\},$$

where $\kappa = |S|$ represents the cardinality of S .

DEFINITION 6.2. A probability vector π^* for S is called *optimal* if $\pi_s^* = 0$ for any $s \notin S^*$.

Observe that if the probability vector is optimal, then the associated state s is in the optimal set S^* with probability one. Furthermore, if $\pi_s^* = 0$ for any $s \notin S^*(a, b)$ then π^* is optimal and the associated state s is again in the optimal set S^* with probability one provided $S^*(a, b) \subset S^*$.

THEOREM 6.1. The probability vector $\pi(M)$ defined in (5.1) converges, as M goes to infinity, to an optimal probability vector π^* . Furthermore

$$\pi_s^* = \begin{cases} 1/|S^*(a, b)|, & \text{if } s \in S^*(a, b), \\ 0, & \text{otherwise,} \end{cases}$$

where $|S^*(a, b)|$ represents the cardinality of $S^*(a, b)$.

The conclusions of the above theorem follow from (5.1) and conclusion (2) of Theorem 3.1. In the next proposition, we establish the monotone property of $\pi_s(M)$ as a function of M .

PROPOSITION 6.1. The following hold:

- (1) For each $s \in S^*(a, b)$, if $M < M'$ then $\pi_s(M) \leq \pi_s(M')$.
- (2) For each $s \notin S^*(a, b)$ there exists an integer M_s such that if $M_s \leq M < M'$ then $\pi_s(M) \geq \pi_s(M')$.

7. Convergence: Analysis for the nonstationary process. In this section we reinstate Assumption A3 and investigate the behavior of the Stochastic Algorithm described in § 4. Because the parameter M_k now varies from iteration to iteration, the algorithm produces a nonstationary Markov chain.

Consider a nonstationary Markov chain with a sequence $\{P(k)\}_{k=1}^\infty$ of state transition probability matrices. We denote by $P(\ell, k)$ the associated state transition probability matrix from iteration ℓ to iteration k where $k \geq \ell \geq 0$. Then

$$(7.1) \quad P(\ell, k) = \begin{cases} \prod_{i=\ell}^{k-1} P(i) & \text{if } k > \ell, \\ I & \text{if } k = \ell. \end{cases}$$

Furthermore, we denote by $x(\ell, k)$ the probability vector of the Markov chain at iteration k when the chain is started at iteration ℓ from the initial probability vector x_0 . Then

$$(7.2) \quad x(\ell, k) = x_0 P(\ell, k), \quad k \geq \ell.$$

Recall that a square matrix $P = (P_{ij})$ is called stochastic if all the entries are positive or zero, $P_{ij} \geq 0$, and the sum of all the entries in each row is one, $\sum_j P_{ij} = 1$.

DEFINITION 7.1. Given a stochastic matrix P , the coefficient of ergodicity is defined to be

$$(7.3) \quad \alpha(P) = \min_{ij} \sum_k \min(P_{ik}, P_{jk}),$$

and the delta coefficient of P is defined to be

$$(7.4) \quad \delta(P) = \frac{1}{2} \max_{ij} \sum_k |P_{ik} - P_{jk}|.$$

It is easy to show [6, p. 143] that

$$(7.5) \quad \alpha(P) = 1 - \delta(P).$$

We are now ready to show that the Markov chain $\{X_k\}$ generated by the Stochastic Algorithm proposed in § 4 is strongly ergodic under certain conditions. Before showing this, we need to obtain a bound on the number of transitions that the Markov chain needs to make before the probability transition matrix has all the elements in at least one column different from zero. For this we first represent the neighborhood system N as a graph by regarding each state $s \in S$ as a node and each neighbor relation $s' \in N(s)$ (hence $s \in N(s')$) as an edge. The length of a path from node s to another node s' is defined to be the number of edges in the path. Then the distance $d(s, s')$ between two nodes s and s' is defined to be the length of a minimum-length path from s to s' . Then the radius of the graph is given by

$$(7.6) \quad r = \min_{s \in S} \max_{s' \in S} d(s, s').$$

Let \hat{s} denote a node at which the above minimum is attained. Then it is called a center of the graph. Furthermore, the radius r represents an upper bound on the number of transitions Markov chain needs to make before the probability transition matrix has all the elements in at least one column, namely the column corresponding to \hat{s} , different from zero.

Now we find the smallest nonzero $R(s, s')$ and the smallest $P(s, a, b)$, as follows:

$$(7.7) \quad \rho = \min_{s \in S} \min_{s' \in N(s)} R(s, s'),$$

$$(7.8) \quad \mu(a, b) = \min_{s \in S} P(s, a, b).$$

It follows from the finiteness of set S and conclusion (2) of Theorem 3.1 that $\rho > 0$ and $0 < \mu(a, b) < 1$.

THEOREM 7.1. *Let a real $c > 0$ satisfy $c \leq 1/r$. Let a real $\sigma > 0$ satisfy $\sigma \geq 1/\mu(a, b)$. Let an integer k_0 satisfy $1 \leq c \log_\sigma(k_0 + 1)$. Let*

$$(7.9) \quad M_k = \text{trunc}[c \log_\sigma(k + k_0 + 1)]$$

for $k = 0, 1, 2, \dots$, where $\text{trunc}[\xi]$ denotes the greatest integer smaller or equal to ξ . Then the Markov chain $\{X_k\}$ generated by the Stochastic Algorithm in § 4 using these M_k is weakly ergodic.

Before showing that the Markov chain $\{X_k\}$ is strongly ergodic, we need to establish the following lemma.

LEMMA 7.1. *The probability vector $\pi(M)$ defined in (5.1) satisfies*

$$\sum_{k=0}^{\infty} \|\pi(M_{k+1}) - \pi(M_k)\| < \infty.$$

THEOREM 7.2. *Let $\{M_k\}$ be as defined in Theorem 7.1. Then the Markov chain $\{X_k\}$ generated by the Stochastic Algorithm in § 4 is strongly ergodic. Furthermore,*

$$(1) \lim_{k \rightarrow \infty} \sup_{x_0} \|x(\ell, k) - \pi^*\| = 0,$$

$$(2) \lim_{k \rightarrow \infty} P[X_k \in S^*(a, b)] = 1,$$

where $x(\ell, k) = x_0 P(\ell, k) = x_0 \prod_{i=\ell}^{k-1} P(M_i)$ and π^* is as defined in Theorem 6.1.

Conclusion (1) in the above theorem states that the probability vector $x(\ell, k)$ for the Markov chain $\{X_k\}$ generated by the Stochastic Algorithm in § 4 converges to π^* as the iteration k goes to infinity. Conclusion (2) states that the probability that the solution estimate X_k is in the optimum set $S^*(a, b)$ for the maximization problem (3.2) converges to one as the iteration k goes to infinity.

In view of Theorem 3.1, $S^*(a, b)$ is contained in the optimum set S^* for the original minimization problem (1.1) under mild conditions. Hence conclusion (2) implies that the probability that the solution estimate X_k is optimum for the original problem converges to one as the iteration k goes to infinity.

8. Rate of convergence. In this section we will show that the probability distribution $x(k)$ of the Markov chain $\{X_k\}$ generated by the Stochastic Algorithm of § 4 converges to the final distribution π^* at a certain speed.

Consider the Markov chain $\{X_k\}$ generated by the Stochastic Algorithm of § 4. The probability distribution $x(k)$ for X_k is expressed below in terms of the transition probability matrix $P(\ell, k)$ from iteration ℓ to k :

$$(8.1) \quad x(k) = x(0)P(0, k),$$

where $x(0)$ represents the initial probability distribution. The final distribution π^* is defined in Theorem 6.1. Hence

$$(8.2) \quad \begin{aligned} \|x(k) - \pi^*\| &= \sum_{s' \in S} \left| \sum_{s \in S} x_s(0) [P_{ss'}(0, k) - \pi_s^*] \right| \\ &\leq \sum_{s \in S} x_s(0) \sum_{s' \in S} |P_{ss'}(0, k) - \pi_s^*|. \end{aligned}$$

Now we develop a bound for the right side of (8.2) in the next lemma. The summation over integers from ℓ to k is to be interpreted as zero when $\ell > k$.

LEMMA 8.1. *For any $s \in S$ and for any integers $\ell \geq 0$ and $k \geq \ell$,*

$$(8.3) \quad \begin{aligned} \sum_{s' \in S} |P_{ss'}(0, k) - \pi_s^*| &\leq 4\delta(P(\ell, k)) \\ &+ \sum_{i=\ell}^{k-1} \|\pi(M_i) - \pi(M_{i+1})\| + \|\pi^* - \pi(M_\ell)\| + \|\pi(M_k) - \pi^*\|. \end{aligned}$$

Observing that the right-hand side of (8.3) is independent of s , we conclude from (8.2) and (8.3) that for any integer $\ell \leq k$,

$$(8.4) \quad \begin{aligned} \|x(k) - \pi^*\| &\leq 4\delta(P(\ell, k)) + \|\pi^* - \pi(M_\ell)\| \\ &+ \sum_{i=\ell}^{k-1} \|\pi(M_i) - \pi(M_{i+1})\| + \|\pi(M_k) - \pi^*\|. \end{aligned}$$

Now we will find a bound on the right-hand side of (8.4). Below, we denote by $P_{\max}(a, b)$ the maximum value for problem (3.2):

$$(8.5) \quad P_{\max}(a, b) = \max \{P(s, a, b) \mid s \in S\}.$$

If the associated optimum set $S^*(a, b)$ were equal to the feasible set S , then the original optimum set S^* would be the same as the feasible set S and both the original and above problems would be trivial. Hence we suppose that $S^*(a, b)$ is not the same as S . Then the second best value

$$(8.6) \quad P_{\text{sec}}(a, b) = \max \{P(s, a, b) \mid s \in S - S^*(a, b)\}$$

is well defined and is less than $P_{\max}(a, b)$.

PROPOSITION 8.1. Suppose that reals c and σ , integer k_0 and a sequence $\{M_k\}$ of positive integers are selected as in Theorem 7.1. Then there exists an integer k^* such that for any integers, $\ell \geq k^*$ and $k \geq \ell$,

$$(8.7) \quad \begin{aligned} & \|\pi^* - \pi(M_\ell)\| + \sum_{i=\ell}^{k-1} \|\pi(M_i) - \pi(M_{i+1})\| + \|\pi(M_k) - \pi^*\| \\ & \leq 4 \frac{|S - S^*(a, b)|}{|S^*(a, b)|} \sigma^\eta (\ell + k_0 + 1)^{-\eta c} \end{aligned}$$

where $\eta = \log_\sigma [P_{\max}(a, b)/P_{\sec}(a, b)] > 0$.

PROPOSITION 8.2. Assume the hypotheses of Proposition 8.1. Then there exists an integer m^* such that for any $m \geq m^*$,

$$(8.8) \quad \delta(P(\ell, mr)) \leq O(1/m^{\hat{t}}),$$

where $\ell = \text{trunc}[\sqrt{m}]r - r$, $\hat{t} = (\rho/r^c)^r/2$, and r and ρ are respectively defined in (7.6) and (7.7).

PROPOSITION 8.3. Assume the hypotheses for Proposition 8.1. Then there exists an integer m^* such that for any $m \geq m^*$,

$$\|\pi^* - \pi(M_\ell)\| + \sum_{i=\ell}^{k-1} \|\pi(M_i) - \pi(M_{i+1})\| + \|\pi(M_k) - \pi^*\| \leq O(1/m^{\bar{t}}),$$

where $k = mr$, $\ell = \text{trunc}[\sqrt{m}]r - r$, and $\bar{t} = \eta c/2$.

THEOREM 8.1. Suppose that reals c and r , integer k_0 , and a sequence $\{M_k\}$ are selected as in Theorem 7.1. Then for a sufficiently large integer m ,

$$\|x(mr) - \pi^*\| \leq O(1/m^t),$$

where $t = \min\{\hat{t}, \bar{t}\} = \min\{(\rho/r^c)^r/2, \eta c/2\} > 0$.

The conclusion for the last theorem follows from (8.4) and Propositions 8.2 and 8.3.

9. Conclusions. In this paper we have proposed a computation method for finding a global solution to the discrete optimization problem in which the objective function needs to be estimated by probabilistic simulation.

The proposed method has a remarkable degree of freedom in the scheme for selecting a candidate Z_k for the next solution estimate X_{k+1} given the current estimate X_k . Indeed some very mild conditions (Assumptions A1–A3) must be satisfied by the neighbor relationship N and the transition probability R for the feasible set S . Hence there is plenty of room for incorporating heuristic ideas in the method, thus making the method very practical and efficient yet theoretically robust, since its theoretical underpinnings are provided by the analysis in this paper.

When designing a practical method based on the algorithm proposed in this paper, one needs to set up a criterion for stopping the infinite process of the proposed method. Ideally, we would like to find some criteria that guarantee that the current estimate X_k is a global optimal solution with α percent confidence. However, such criteria appear to be difficult to find. Therefore, for the moment, we suggest the usual criteria: stop the method when the parameter M_k , the iteration counter k , the number of simulations, or the number of objective value samples exceeds a preselected fixed number.

The proposed algorithm is related to, but different from, the simulated annealing algorithm. While the objective value in simulated annealing is assumed to be exactly computable, the proposed algorithm accepts samples of the objective value generated by probabilistic simulation. The two algorithms are not equivalent even when the

parameter M_k goes to infinity, because the proposed algorithm does not rely on the sample mean to estimate the objective value.

Appendix.

Proof of Theorem 3.1. It follows from the finiteness of S that there exist reals c_1 and c_2 such that

$$(A1) \quad g(s) = E[H(s)] \in [c_1, c_2], \quad \forall s \in S.$$

It also follows that

$$(A2) \quad \varepsilon \triangleq \min \{|g(s) - g(s')| \mid (s, s') \in G\} > 0,$$

where $G = \{(s, s') \in S \times S \mid g(s) \neq g(s')\}$.

Let $D_s: \mathbb{R}^2 \rightarrow \mathbb{R}$ be defined by

$$D_s(a, b) = \int_{h < a}^{h > b} h F_{H(s)}(dh).$$

Not it follows from (2.3) that there exist reals c_3 and c_4 such that if $a \leq c_3$ and $b \geq c_4$ then

$$(A3) \quad |D_s(a, b)| \leq \varepsilon/4, \quad \forall s \in S.$$

Let

$$(A4) \quad V_{\max} = \max \{E[H(s)^2] \mid s \in S\}.$$

Let $c_5 = 8V_{\max}/\varepsilon$. Let $\bar{a} = \min\{c_1, c_3, -c_5\}$ and $\bar{b} = \max\{c_2, c_4, c_5\}$. Then $\bar{a} \leq -c_5 < 0 < c_5 \leq \bar{b}$.

Suppose that $a < \bar{a}$, $\bar{b} < b$ and $s \in S$. Then

$$\begin{aligned} P(s, a, b) &= P\{[H(s) \leq \Theta(a, b)] \cap [a \leq H(s) \leq b]\} \\ &\quad + P\{[H(s) \leq \Theta(a, b)] \cap [H(s) < a]\} \\ &\quad + P\{[H(s) \leq \Theta(a, b)] \cap [H(s) > b]\} \\ &= \int_a^b \int_h^b F_{\Theta}(d\theta) F_{H(s)}(dh) + P[H(s) < a] \\ &= \int_a^b \frac{b-h}{b-a} F_{H(s)}(dh) + P[H(s) < a] \\ &= \frac{b}{b-a} \int_a^b F_{H(s)}(dh) - \frac{1}{b-a} \int_a^b h F_{H(s)}(dh) + P[H(s) < a] \\ &= \frac{b}{b-a} \{1 - P[H(s) > b] - P[H(s) < a]\} \\ &\quad - \frac{1}{b-a} \{E[H(s)] - D_s(a, b)\} + P[H(s) < a] \\ &= \frac{1}{b-a} \{[b - g(s)] - bP[H(s) > b] - aP[H(s) < a] + D_s(a, b)\}. \end{aligned}$$

Hence, recalling $a < 0 < b$, we obtain

$$\begin{aligned}
 & P(s, a, b) - P(s', a, b) \\
 &= \frac{1}{b-a} [\{g(s') - g(s)\} + b\{P[H(s') > b] - P[H(s) > b]\} \\
 &\quad + a\{P[H(s') < a] - P[H(s) < a]\} + \{D_s(a, b) - D_{s'}(a, b)\}] \\
 &\geq \frac{1}{b-a} \{g(s') - g(s) - bP[H(s) > b] + aP[H(s') < a] \\
 (A5) \quad &\quad + D_s(a, b) - D_{s'}(a, b)\} \\
 &\geq \frac{1}{b-a} \{g(s') - g(s) - bP[|H(s)| \geq b] + aP[|H(s')| \geq |a|] \\
 &\quad - |D_s(a, b)| - |D_{s'}(a, b)|\} \\
 &\geq \frac{1}{b-a} \{g(s') - g(s) - \frac{1}{b} E[H(s)^2] + \frac{1}{a} E[H(s')^2] \\
 &\quad - |D_s(a, b)| - |D_{s'}(a, b)|\},
 \end{aligned}$$

where the last inequality follows from the Markov inequality

$$P[|H(s)| \geq c] \leq E[H(s)^2]/c^2.$$

If $g(s) < g(s')$, then it follows from (A2)-(A5) that

$$\begin{aligned}
 P(s, a, b) - P(s', a, b) &\geq \frac{1}{b-a} \left\{ \varepsilon - \frac{1}{b} V_{\max} + \frac{1}{a} V_{\max} - \varepsilon/2 \right\} \\
 &\geq \varepsilon/4(b-a) > 0.
 \end{aligned}$$

This establishes conclusion (1).

It now follows from (A1) that $g(s) = E[H(s)] \in (a, b)$, for all $s \in S$. Hence $P[H(s) \leq g(s)] > 0$ and $P[g(s) \leq \Theta(a, b)] > 0$. Noting

$$\{H(s) \leq g(s)\} \cap \{g(s) \leq \Theta(a, b)\} \subset \{H(s) \leq \Theta(a, b)\},$$

we conclude

$$P[H(s) \leq \Theta(a, b)] \geq P[H(s) \leq g(s)]P[g(s) \leq \Theta(a, b)] > 0.$$

Similarly, we conclude

$$P[H(s) > \Theta(a, b)] \geq P[H(s) \geq g(s)]P[g(s) > \Theta(a, b)] > 0.$$

Hence $P[H(s) \leq \Theta(a, b)] = 1 - P[H(s) > \Theta(a, b)] < 1$. These inequalities imply conclusion (2).

Consider the maximization problem (3.2). Since S is finite, the maximum exists and $S^*(a, b) \neq \emptyset$. Let $s \in S^*(a, b)$. Then $P(s, a, b) \geq P(s', a, b)$ for any $s' \in S$. It now follows from conclusion (1) that $g(s) \leq g(s')$ for any $s' \in S$. Hence $s \in S^*$. In other words, $S^*(a, b) \subset S^*$. \square

Proof of Theorem 3.2. Let $\bar{a} = \min \{a(s) | s \in S\}$ and $\bar{b} = \max \{b(s) | s \in S\}$. Let $a < \bar{a}$ and $b > \bar{b}$. Now for each $s \in S$,

$$\begin{aligned}
 P(s, a, b) &= P[H(s) \leq \Theta(a, b)] \\
 &= \int_{a(s)}^{b(s)} \int_h^b F_\theta(d\theta) F_{H(s)}(dh) \\
 &= \int_{a(s)}^{b(s)} \frac{(b-h)}{b-a} F_{H(s)}(dh)
 \end{aligned}$$

$$\begin{aligned}
&= \frac{b}{b-a} \int_{a(s)}^{b(s)} F_{H(s)}(dh) - \frac{1}{b-a} \int_{a(s)}^{b(s)} h F_{H(s)}(dh) \\
&= \frac{1}{b-a} \{b - E[H(s)]\} = \frac{1}{b-a} \{b - g(s)\}.
\end{aligned}$$

Hence

$$P(s, a, b) - P(s', a, b) = \frac{1}{b-a} \{g(s') - g(s)\}.$$

Therefore $g(s) \leq g(s')$ if and only if $P(s, a, b) \geq P(s', a, b)$. Hence $S^*(a, b) = S^*$. \square

The following proposition is needed in the proof of Theorem 5.1.

PROPOSITION A.1. *The stationary Markov chain $\{X_k\}$ defined by (4.1) with $M_k = M$ is irreducible and aperiodic.*

Proof. The irreducibility follows from Assumption A1, Definition 4.2, (4.1), and conclusion (2) in Theorem 3.1. The aperiodicity follows from (4.1) and conclusion (2) in Theorem 3.1. \square

Proof of Theorem 5.1. It follows from Theorem 3.1 and (5.1) that $\pi_s(M) > 0$ for each $s \in S$, and that

$$(A6) \quad \|\pi(M)\| = \sum_{s \in S} \pi_s(M) = 1$$

where $\|\cdot\|$ denotes the ℓ_1 -norm.

It follows from Assumption A2, (5.1), and (4.1) that for every pair (s, s') of neighbors, $s' \in N(s)$,

$$\frac{\pi_s(M)}{\pi_{s'}(M)} = \frac{\{P(s, a, b)\}^M}{\{P(s', a, b)\}^M} = \frac{R(s', s)\{P(s, a, b)\}^M}{R(s, s')\{P(s', a, b)\}^M} = \frac{P_{s's}(M)}{P_{ss'}(M)}$$

and that

$$(A7) \quad \pi_s(M) P_{ss'}(M) = \pi_{s'}(M) P_{s's}(M).$$

Observe that this equation trivially holds when $s = s'$ and even when s and s' are not neighbors to each other, in which case $P_{ss'}(M) = P_{s's}(M) = 0$. In other words, (A7) is valid for any s and s' in S . It now follows from (A7) that

$$(A8) \quad \pi_{s'}(M) = \sum_{s \in S} \pi_s(M) P_{s's}(M) = \sum_{s \in S} \pi_s(M) P_{ss'}(M).$$

By virtue of [6, Thms. III.2.2 and III.2.1], we conclude from (A6) and (A8) that $\{\pi_s(M) | s \in S\}$ is indeed the stationary probability distribution. \square

Proof of Proposition 6.1. Consider $\pi_s(M)$ as a function of a real variable M . Then π_s is differentiable with respect to M . Noting $d\alpha^M/dM = \alpha^M \ln \alpha$ and working out the details, we arrive at

$$\frac{d\pi_s(M)}{dM} = \left\{ \sum_{s' \in S} \left[\frac{P(s', a, b)}{P(s, a, b)} \right]^M \ln \left[\frac{P(s, a, b)}{P(s', a, b)} \right] \right\} \{\pi_s(M)\}^2$$

Suppose that $s \in S^*(a, b)$. Then $\ln [P(s, a, b)/P(s', a, b)] = 0$ for all $s' \in S^*(a, b)$ so that

$$\frac{d\pi_s(M)}{dM} = \left\{ \sum_{s' \in S - S^*(a, b)} \left[\frac{P(s', a, b)}{P(s, a, b)} \right]^M \ln \left[\frac{P(s, a, b)}{P(s', a, b)} \right] \right\} \{\pi_s(M)\}^2.$$

Furthermore, for any $s' \in S - S^*(a, b)$, $\ln [P(s, a, b)/P(s', a, b)] > 0$, so that $d\pi_s(M)/dM > 0$ for any $M > 0$. This implies conclusion (1).

Suppose that $s \notin S^*(a, b)$. Then

$$\begin{aligned} \frac{d\pi_s(M)}{dM} = & \left\{ \sum_{P(s,a,b) > P(s',a,b)} \left[\frac{P(s', a, b)}{P(s, a, b)} \right]^M \ln \left[\frac{P(s, a, b)}{P(s', a, b)} \right] \right. \\ & \left. - \sum_{P(s,a,b) < P(s',a,b)} \left[\frac{P(s', a, b)}{P(s, a, b)} \right]^M \ln \left[\frac{P(s', a, b)}{P(s, a, b)} \right] \right\} \{\pi_s(M)\}^2. \end{aligned}$$

Observe that the first term monotonically decreases to zero while the second term monotonically increases to infinity in magnitude as M goes to infinity. Hence $d\pi_s(M)/dM$ monotonically decreases to negative infinity. In other words, there exists a real M_s such that $d\pi_s(M)/dM < 0$ for any $M \geq M_s$. This implies conclusion (2). \square

Proof of Theorem 7.1. Let c, σ, k_0 be chosen as above. Then $\mu(a, b) \geq 1/\sigma$. Now if s and s' are neighbors to each other in S , then it follows from (4.1), (7.7), and (7.8) that

$$P_{ss'}(M_k) \geq \rho \left(\frac{1}{\sigma} \right)^{M_k}.$$

It also follows from (4.1) that $P_{ss}(M_k)$ monotonically increases as k increases for each $s \in S$. Since $(1/\sigma)^{M_k}$ monotonically decreases as k increases, there exists an integer k^* such that

$$P_{ss}(M_k) \geq \rho \left(\frac{1}{\sigma} \right)^{M_k}, \quad \forall k \geq k^*, \quad \forall s \in S.$$

Let $P(\ell, k)$ denote the transition probability matrix from iteration ℓ to iteration $k > \ell$. Then

$$P(\ell, k) = \prod_{i=\ell}^{k-1} P(M_i).$$

Recalling that after r iterations there is at least one path in the graph from the center \hat{s} to any $s \in S$, we observe that the entries in the column of $P(\ell, k)$ corresponding to the center \hat{s} satisfy for any $k \geq k^* + r$

$$(A9) \quad P_{s\hat{s}}(k-r, k) \geq \prod_{i=k-r}^{k-1} \left\{ \rho \left(\frac{1}{\sigma} \right)^{M_i} \right\} \geq \rho^r \left(\frac{1}{\sigma} \right)^{rM_{k-1}}, \quad \forall s \in S.$$

Let k be a multiple of r , i.e., $k = mr$. Then it follows from (7.3) and (A9) that the coefficient of ergodicity for $P(mr-r, mr)$ satisfies

$$\begin{aligned} \alpha(P(mr-r, mr)) &= \min_{s, s'} \sum_{s'' \in S} \min \{P_{ss''}(mr-r, mr), P_{s's''}(mr-r, mr)\} \\ &\geq \min_{s, s'} \min \{P_{s\hat{s}}(mr-r, mr), P_{s'\hat{s}}(mr-r, mr)\} \\ &\geq \rho^r \left(\frac{1}{\sigma} \right)^{rM_{mr-1}}, \quad \forall m \geq \frac{1}{r}(k^* + r). \end{aligned}$$

Let m^* be the smallest integer satisfying $m^* \geq (1/r)(k^* + r)$. It now follows from (7.9)

that $M_k \leq c \log_\sigma(k + k_0 + 1)$ and

$$\left(\frac{1}{\sigma}\right)^{M_k} \geq \frac{1}{(k + k_0 + 1)^c}$$

so that

$$\left(\frac{1}{\sigma}\right)^{rM_{mr-1}} \geq \frac{1}{(mr + k_0)^{cr}}.$$

Therefore

$$\sum_{m=m^*}^{\infty} \alpha(P(mr - r, mr)) \geq \sum_{m=m^*}^{\infty} \rho^r \frac{1}{(mr + k_0)^{cr}}.$$

The right-hand side goes to infinity since $0 < cr \leq 1$. It now follows from [6, Thm. V.3.2] that the Markov chain $\{X_k\}$ is weakly ergodic. \square

Proof of Lemma 7.1. It follows from Proposition 6.1 that there exists an integer k^* such that for any $k \geq k^*$,

$$\begin{aligned} \pi_s(M_{k+1}) &\geq \pi_s(M_k), \quad \forall s \in S^*(a, b), \\ \pi_s(M_{k+1}) &\leq \pi_s(M_k), \quad \forall s \notin S^*(a, b). \end{aligned}$$

Hence, for any $k \geq k^*$,

$$\|\pi(M_{k+1}) - \pi(M_k)\| = \sum_{s \in S^*(a, b)} [\pi_s(M_{k+1}) - \pi_s(M_k)] - \sum_{s \notin S^*(a, b)} [\pi_s(M_{k+1}) - \pi_s(M_k)].$$

Noting

$$\sum_{s \in S^*(a, b)} \pi_s(M_k) + \sum_{s \notin S^*(a, b)} \pi_s(M_k) = \|\pi(M_k)\| = 1,$$

we conclude that for any $k \geq k^*$

$$\|\pi(M_{k+1}) - \pi(M_k)\| = 2 \sum_{s \in S^*(a, b)} [\pi_s(M_{k+1}) - \pi_s(M_k)].$$

Hence, for any $\ell \geq k^*$,

$$\begin{aligned} \sum_{k=k^*}^{\ell} \|\pi(M_{k+1}) - \pi(M_k)\| &= 2 \sum_{s \in S^*(a, b)} [\pi_s(M_{\ell+1}) - \pi_s(M_{k^*})] \\ &\leq 2 \sum_{s \in S^*(a, b)} \pi_s(M_{\ell+1}) \leq 2. \end{aligned} \quad \square$$

Proof of Theorem 7.2. It follows from Theorem 7.1 that the Markov chain is weakly ergodic. It now follows from (5.1), Theorem 5.1, Lemma 7.1, Theorem 6.1, and [6, Thm. V.4.3] that the Markov chain is strongly ergodic and that conclusions (1) and (2) hold. \square

Proof of Lemma 8.1. Let $s \in S$ and $0 \leq \ell \leq k$ be fixed

$$\begin{aligned} &\sum_{s' \in S} |P_{ss'}(0, k) - \pi_{s'}^*| \\ &= \sum_{s' \in S} \left| \sum_{\bar{s} \in S} P_{s\bar{s}}(0, \ell) P_{\bar{s}s'}(\ell, k) - \pi_{s'}^* \right| \\ (A10) \quad &= \sum_{s' \in S} \left| \sum_{\bar{s} \in S} [P_{s\bar{s}}(0, \ell) - \pi_{\bar{s}}^*] P_{\bar{s}s'}(\ell, k) + \sum_{\bar{s} \in S} \pi_{\bar{s}}^* P_{\bar{s}s'}(\ell, k) - \pi_{s'}^* \right| \\ &\leq \sum_{s' \in S} \left| \sum_{\bar{s} \in S} [P_{s\bar{s}}(0, \ell) - \pi_{\bar{s}}^*] P_{\bar{s}s'}(\ell, k) \right| + \sum_{s' \in S} \left| \sum_{\bar{s} \in S} \pi_{\bar{s}}^* P_{\bar{s}s'}(\ell, k) - \pi_{s'}^* \right|. \end{aligned}$$

Consider the first term of the last expression of (A10).

$$\begin{aligned}
 \sum_{\bar{s} \in S} [P_{s\bar{s}}(0, \ell) - \pi_{\bar{s}}^*] &= \sum_{\bar{s} \in S} P_{s\bar{s}}(0, \ell) - \sum_{\bar{s} \in S} \pi_{\bar{s}}^* = 1 - 1 = 0, \\
 \sum_{s' \in S} \left| \sum_{\bar{s} \in S} [P_{s\bar{s}}(0, \ell) - \pi_{\bar{s}}^*] P_{\bar{s}s'}(\ell, k) \right| \\
 &= \max_{s'' \in S} \sum_{s' \in S} \left| \sum_{\bar{s} \in S} [P_{s\bar{s}}(0, \ell) - \pi_{\bar{s}}^*] [P_{s''s'}(\ell, k) - P_{\bar{s}s'}(\ell, k)] \right| \\
 &\leq \sum_{\bar{s} \in S} \left| P_{s\bar{s}}(0, \ell) - \pi_{\bar{s}}^* \right| \max_{s'' \in S} \sum_{s' \in S} |P_{s''s'}(\ell, k) - P_{\bar{s}s'}(\ell, k)| \\
 &\leq \sum_{\bar{s} \in S} |P_{s\bar{s}}(0, \ell) - \pi_{\bar{s}}^*| \max_{\bar{s}, s'' \in S} \sum_{s' \in S} |P_{s''s'}(\ell, k) - P_{\bar{s}s'}(\ell, k)| \\
 &\leq 4\delta(P(\ell, k)),
 \end{aligned}
 \tag{A11}$$

where the last inequality follows from (7.4) and

$$\sum_{\bar{s} \in S} |P_{s\bar{s}}(0, l) - \pi_{\bar{s}}^*| \leq 2.$$

Now consider the second term of the last expression of (A10). To find a bound on that term, we take advantage of the matrix norm induced by the vector ℓ_1 -norm $\| \cdot \|$ as follows:

$$\|A\| = \max_i \sum_j |a_{ij}|,$$

where $A = (a_{ij})$. Observe that $\|xA\| \leq \|x\| \|A\|$.

Let Q be the $\kappa \times \kappa$ matrix whose rows are all π^* . Let Q_k be the $\kappa \times \kappa$ matrix whose rows are all $\pi(M_k)$ for $k = 0, 1, 2, \dots$. Then

$$\begin{aligned}
 \sum_{s' \in S} \left| \sum_{\bar{s} \in S} \pi_{\bar{s}}^* P_{\bar{s}s'}(\ell, k) - \pi_{s'}^* \right| \\
 &= \|QP(\ell, k) - Q\| \\
 &\leq \|QP(\ell, k) - Q_\ell P(\ell, k)\| + \|Q_\ell P(\ell, k) - Q_k\| + \|Q_k - Q\|.
 \end{aligned}
 \tag{A12}$$

Since $\|P(\ell, k)\| = 1$, we find

$$\|QP(\ell, k) - Q_\ell P(\ell, k)\| = \|Q - Q_\ell\| \|P(\ell, k)\| = \|Q - Q_\ell\|.$$

It follows from (5.2) that

$$\pi(M_i)P(M_i) = \pi(M_i), \quad \forall i, \tag{A14}$$

$$Q_i P(i, i+1) = Q_i P(M_i) = Q_i, \quad \forall i. \tag{A15}$$

Suppose $k > \ell$. Then it follows from (A15) that

$$\begin{aligned}
 Q_\ell P(\ell, k) &= Q_\ell P(\ell, \ell+1)P(\ell+1, k) = Q_\ell P(\ell+1, k) \\
 &= (Q_\ell - Q_{\ell+1})P(\ell+1, k) + Q_{\ell+1}P(\ell+1, k).
 \end{aligned}$$

We may continue this process and obtain

$$Q_\ell P(\ell, k) = \sum_{i=\ell}^{k-1} (Q_i - Q_{i+1})P(i+1, k) + Q_k.$$

Observe that this equality holds even when $k = \ell$. It now follows that

$$\|Q_\ell P(\ell, k) - Q_k\| = \left\| \sum_{i=\ell}^{k-1} (Q_i - Q_{i+1})P(i+1, k) \right\| \leq \sum_{i=\ell}^{k-1} \|Q_i - Q_{i+1}\|. \tag{A16}$$

Collecting (A12), (A13), and (A16), we arrive at

$$\begin{aligned}
 (A17) \quad & \sum_{s' \in S} \left| \sum_{\bar{s} \in S} \pi_{\bar{s}}^* P_{\bar{s}s'}(\ell, k) - \pi_{s'}^* \right| \\
 & \leq \|Q - Q_\ell\| + \sum_{i=\ell}^{k-1} \|Q_i - Q_{i+1}\| + \|Q_k - Q\| \\
 & = \|\pi^* - \pi(M_\ell)\| + \sum_{i=\ell}^{k-1} \|\pi(M_i) - \pi(M_{i+1})\| + \|\pi(M_k) - \pi^*\|.
 \end{aligned}$$

Now (8.3) follows from (A10), (A11), and (A17). \square

Proof of Proposition 8.1. It follows from Proposition 6.1 that there exists an integer \hat{k} such that for any $k \geq \hat{k}$,

$$\pi_s(M_{k+1}) \geq \pi_s(M_k), \quad \forall s \in S^*(a, b),$$

$$\pi_s(M_{k+1}) \leq \pi_s(M_k), \quad \forall s \notin S^*(a, b).$$

Hence, recalling $\pi^* = \lim_{k \rightarrow \infty} \pi_s(M_k)$, we obtain for any integers $\ell \geq \hat{k}$ and $k \geq \ell$.

$$\begin{aligned}
 & \|\pi^* - \pi(M_\ell)\| + \sum_{i=\ell}^{k-1} \|\pi(M_i) - \pi(M_{i+1})\| + \|\pi(M_k) - \pi^*\| \\
 &= \sum_{s \in S^*(a, b)} \left\{ |\pi_s^* - \pi_s(M_\ell)| + \sum_{i=\ell}^{k-1} |\pi_s(M_{i+1}) - \pi_s(M_i)| + |\pi_s^* - \pi_s(M_k)| \right\} \\
 & \quad + \sum_{s \notin S^*(a, b)} \left\{ |\pi_s(M_\ell) - \pi_s^*| + \sum_{i=\ell}^{k-1} |\pi_s(M_i) - \pi_s(M_{i+1})| + |\pi_s(M_k) - \pi_s^*| \right\} \\
 &= 2 \left\{ \sum_{s \in S^*(a, b)} [\pi_s^* - \pi_s(M_\ell)] + \sum_{s \notin S^*(a, b)} [\pi_s(M_\ell) - \pi_s^*] \right\} \\
 &\leq 2 \left\{ 1 - \sum_{s \in S^*(a, b)} \pi_s(M_\ell) + \sum_{s \notin S^*(a, b)} \pi_s(M_\ell) \right\} \\
 &= 4 \left\{ 1 - \sum_{s \in S^*(a, b)} \pi_s(M_\ell) \right\}.
 \end{aligned}$$

Now it follows from (5.1) that

$$\begin{aligned}
 & 1 - \sum_{s \in S^*(a, b)} \pi_s(M_\ell) \\
 &= 1 - \sum_{s \in S^*(a, b)} \frac{\{P(s, a, b)\}^{M_\ell}}{\sum_{s' \in S} \{P(s', a, b)\}^{M_\ell}} \\
 &= \frac{\sum_{s \notin S^*(a, b)} \{P(s, a, b)\}^{M_\ell}}{\sum_{s \in S^*(a, b)} \{P(s, a, b)\}^{M_\ell} + \sum_{s \notin S^*(a, b)} \{P(s, a, b)\}^{M_\ell}} \\
 &\leq \frac{\sum_{s \notin S^*(a, b)} \{P(s, a, b)\}^{M_\ell}}{\sum_{s \in S^*(a, b)} \{P(s, a, b)\}^{M_\ell}} \\
 &= \frac{\sum_{s \notin S^*(a, b)} \{P(s, a, b)\}^{M_\ell}}{|S^*(a, b)| \{P_{\max}(a, b)\}^{M_\ell}}
 \end{aligned}$$

$$\begin{aligned}
&\leq \frac{1}{|S^*(a, b)|} \sum_{s \notin S^*(a, b)} \left\{ \frac{P_{\text{sec}}(a, b)}{P_{\text{max}}(a, b)} \right\}^{M_\ell} \\
&= \frac{|S - S^*(a, b)|}{|S^*(a, b)|} \left(\frac{1}{\sigma} \right)^{\eta M_\ell}.
\end{aligned}$$

It now follows from the selection (7.9) for M_ℓ that

$$\left(\frac{1}{\sigma} \right)^{M_\ell} \leq \frac{\sigma}{(\ell + k_0 + 1)^c}.$$

Hence we arrive at (8.7). \square

Proof of Proposition 8.2. According to the proof of Theorem 7.1, there exists an m^* such that for any $m \geq m^*$,

$$\alpha(P((m-1)r, mr)) = 1 - \delta(P((m-1)r, mr)) \leq \rho^r / (mr + k_0)^{cr}.$$

Suppose that k is an integer multiple of integer r , i.e., $k = mr$, so that $\text{trunc}[\sqrt{m}] \geq m^*$. Let $\ell = \text{trunc}[\sqrt{m}]r - r$ and $\bar{m} = \text{trunc}[\sqrt{m}]$. Then it follows from [6, Lemma V.2.3, p. 145] that

$$\begin{aligned}
\delta(P(\ell, k)) &= \delta(P(\text{trunc}[\sqrt{m}]r - r, mr)) \leq \prod_{i=\bar{m}}^m \delta(P((i-1)r, ir)) \\
&\leq \prod_{i=\bar{m}}^m \left\{ 1 - \frac{\rho^r}{(ir + k_0)^{cr}} \right\} \\
&= \prod_{i=\bar{m}}^m \left\{ 1 - \frac{(\rho/r^c)^r}{[i + (k_0/r)]^{rc}} \right\} \\
&\leq \prod_{i=\bar{m}}^m \exp \left\{ -\frac{2\hat{t}}{[i + (k_0/r)]^{rc}} \right\},
\end{aligned}$$

where the last inequality is derived from $1 + y \leq e^y$. Since $rc \leq 1$, we obtain

$$\delta(P(\ell, k)) \leq \exp \left\{ -2\hat{t} \sum_{i=\bar{m}}^m \frac{1}{[i + (k_0/r)]} \right\} \leq \left\{ \frac{\sqrt{m} + (k_0/r)}{m + k_0/r} \right\}^{2\hat{t}},$$

where the last inequality follows from the fact that

$$\sum_{i=n_1}^{n_2} \frac{1}{i} \geq \int_{n_1}^{n_2} \frac{1}{x} dx = \ln \left(\frac{n_2}{n_1} \right)$$

for any positive integers n_1 and $n_2 \geq n_1$.

The conclusion now follows from the above inequality. \square

Proof of Proposition 8.3. Suppose $k = mr$ and $\ell = \text{trunc}[\sqrt{m}]r - r$ are sufficiently large so that (8.7) holds. Then

$$\begin{aligned}
&\|\pi^* - \pi(M_\ell)\| + \sum_{i=\ell}^{k-1} \|\pi(M_i) - \pi(M_{i+1})\| + \|\pi(M_k) - \pi^*\| \\
&\leq 4 \frac{|S - S^*(a, b)|}{|S^*(a, b)|} \sigma^\eta \frac{1}{(\ell + k_0 + 1)^{\eta c}} \\
&\leq 4 \frac{|S - S^*(a, b)|}{|S^*(a, b)|} \sigma^\eta \frac{(1/r)^{\eta c}}{\left(\sqrt{m} + \frac{k_0 + 1}{r} - 1 \right)^{\eta c}} = O(1/m^{\bar{t}})
\end{aligned}$$

for a large m . \square

REFERENCES

- [1] P. W. GLYNN, *Optimization of stochastic systems*, in Proc. 1986 Winter Simulation Conference, 1986, pp. 52–59.
- [2] P. W. GLYNN AND J. L. SANDERS, *Monte Carlo optimization of stochastic systems: Two new approaches*, in Proc. 1986 ASME Computers in Engineering Conference, 1986, pp. 219–223.
- [3] D. GROSS AND C. M. HARRIS, *Fundamentals of Queueing Theory*, John Wiley, New York, 1985.
- [4] B. HAJEK, *A tutorial survey of theory and applications of simulated annealing*, in Proc. 24th IEEE Conf. on Decision and Control, 1985, pp. 755–760.
- [5] Y. C. HO, M. A. EYLER, AND T. T. CHIEN, *A gradient technique for general buffer storage design in a serial production line*, Internat. J. Production Res., 17 (1979), pp. 557–580.
- [6] D. L. ISAACSON AND R. W. MADSEN, *Markov Chains Theory and Applications*, John Wiley, New York, 1976.
- [7] R. KARP, *Probabilistic analysis of partitioning algorithms for the traveling salesman problem on the plane*, Math. Oper. Res., 2 (1977), pp. 209–224.
- [8] S. KIRKPATRICK, C. D. GELATT, JR, AND M. P. VECCHI, *Optimization by simulated annealing*, Science, 220 (1983), pp. 671–680.
- [9] E. L. LAWLER AND D. E. WOOD, *Branch and bound methods: A survey*, Oper. Res., 14 (1966), pp. 699–719.
- [10] S. LIN, *Heuristic programming as an aid to network design*, Networks, 5 (1975), pp. 33–43.
- [11] D. MITRA, F. ROMEO, AND A. SANGIOVANNI-VINCENTELLI, *Convergence and finite-time behavior of simulated annealing*, in Proc. 24th IEEE Conf. Decision and Control, 1985, pp. 761–767.
- [12] C. H. PAPADIMITRIOU AND K. STEIGLITZ, *Combinatorial Optimization Algorithms and Complexity*, Prentice-Hall, Englewood Cliffs, NJ, 1982.
- [13] J. G. SHANTHIKUMAR AND D. D. YAO, *Strong stochastic convexity and its applications in parametric optimization of queueing systems*, in Proc. 27th IEEE Conf. on Decision and Control, 1988, pp. 657–661.

INVESTMENT-CONSUMPTION MODELS WITH TRANSACTION FEES AND MARKOV-CHAIN PARAMETERS*

THALEIA ZARIPHOUPOULOU†

Abstract. This paper considers an infinite horizon investment-consumption model in which a single agent consumes and distributes his wealth in two assets, a bond and a stock. The problem of maximization of the total utility from consumption is treated. State (amount allocated in assets) and control (consumption, rates of trading) constraints are present. It is shown that the value function is the unique viscosity solution of a system of variational inequalities with gradient constraints.

Key words. viscosity solutions, state constraints, variational inequalities, singular controls

AMS(MOS) subject classification. 90A35

Introduction. In this paper we examine a general investment and consumption decision problem for a single agent. The investor consumes at a nonnegative rate and he distributes his current wealth between two assets continuously in time. One asset is a bond, i.e., a riskless security with instantaneous rate of return r . The other asset is a stock, whose rate of return z_t is a continuous time Markov chain. In our version of the model the investor *cannot borrow* money to finance his investment in bond and he *cannot short-sell* the stock. In other words, the amount of money allocated in bond and stock must stay nonnegative.

When the investor makes a transaction, he pays transaction fees which are assumed to be proportional to the amount transacted. More specifically, let x_t and y_t be the investor's holdings in the riskless and the risky security prior to a transaction at time t . If the investor increases (or decreases) the amount invested in the risky asset to $y_t + h_t$ (or $y_t - h_t$), the holding of the riskless asset decreases (increases) to $x_t - h_t - \lambda h_t$ (or $x_t + h_t - \mu h_t$). The numbers λ and μ are assumed to be nonnegative and one of them must always be positive. The control objective is to maximize, in an infinite horizon, the expected discounted utility which comes only from consumption. Due to the presence of the transaction fees, this is a singular control problem.

The paper is organized as follows. Section 1 is devoted to the description of the model and its history; the two main theorems are also stated here. Section 2 contains preliminaries about the value function. In § 3 we approximate the problem by using absolutely continuous controls. Finally, in § 4 we show that the value function is the unique constrained viscosity solution of a system of Variational Inequalities with gradient constraints.

1. We consider a market with two assets: a *bond* and a *stock*. The price P_t^0 of the bond is given by

$$(1.1) \quad \begin{aligned} dP_t^0 &= rP_t^0 dt, \\ P_0^0 &= p_0, \end{aligned}$$

where $r > 0$. The price P_t of the stock satisfies

$$(1.2) \quad \begin{aligned} dP_t &= z(t)P_t dt, \\ P_0 &= p. \end{aligned}$$

* Received by the editors November 20, 1989; accepted for publication (in revised form) December 21, 1990.

† Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, Massachusetts 01609. This work was partially supported by National Science Foundation Research planning grant 530177.

The rate of return z is a *finite state continuous time Markov chain*, defined on some underlying probability space (Ω, F, P) with jumping rate $q_{zz'}$ from state z to state z' . The state space is denoted by Z . The associated generator \mathcal{L} of the Markov chain has the form

$$\mathcal{L}v(z) = \sum_{z' \neq z} q_{zz'}[v(z') - v(z)].$$

Let $K = \max_{z \in Z} z$. A natural assumption is $K \geq r$. The amount of wealth x_t and y_t , invested at time t in bond and stock respectively, are the state variables and they evolve (see [17]) according to the equations

$$\begin{aligned} dx_t &= (rx_t - C_t) dt - (1 + \lambda) dM_t + (1 - \mu) dN_t, \\ dy_t &= z(t)y_t dt + dM_t - dN_t, \\ x_0 &= x, y_0 = y, z(0) = z. \end{aligned} \quad (1.3)$$

For simplicity we assume here that all financial charges are paid from the holdings in bond. The investor cannot borrow money or short-sell the stock. The control processes are the *consumption rate* C_t and the processes M_t and N_t which represent the *cumulative purchases* and *sales of stock* respectively. The controls (C_t, M_t, N_t) are *admissible* if

(i) C_t is F_t -measurable, where $F_t = \sigma(z_s : 0 \leq s \leq t)$, $C_t \geq 0$ almost everywhere for all $t \geq 0$, and $E \int_0^\infty e^{-rs} C_s ds < +\infty$.

(ii) M_t, N_t are F_t -measurable, right continuous, and nondecreasing processes.

(iii) $x_t \geq 0, y_t \geq 0$ almost everywhere for all $t \geq 0$, where x_t, y_t are the trajectories given by the state equation (1.3) using the controls (C_t, M_t, N_t) . We denote by A the set of admissible controls.

The total expected discounted utility J coming from consumption is given by

$$J(x, y, z, C, M, N) = E \int_0^{+\infty} e^{-\beta t} U(C_t) dt$$

with $(C, M, N) \in A$ and $z(0) = z$, where the utility function $U : [0, +\infty) \rightarrow [0, +\infty)$ is assumed to have the following properties:

U is strictly increasing, bounded, concave, C^1 function,

and

$$U(0) = 0, \lim_{c \rightarrow 0} U'(c) = +\infty, \lim_{c \rightarrow \infty} U'(c) = 0.$$

The discount factor $\beta > 0$ weights consumption now versus consumption later, large β denoting instant gratification. Note that the controls M and N are acting implicitly through the constraint (iii).

The value function u is given by

$$u(x, y, z) = \sup_A E \int_0^{+\infty} e^{-\beta t} U(C_t) dt.$$

Our goal is to derive the Bellman equation associated with this singular control problem and to characterize u as its unique solution. It turns out that the Bellman equation here is a system of variational inequalities.

We now state one of the main results. (For the definition of constrained viscosity solution, see Definition 3.1.)

THEOREM. *The value function u is the unique constrained viscosity solution of*

$$(1.4) \quad \begin{aligned} & \min [(1+\lambda)u_x - u_y, -(1-\mu)u_x + u_y, \\ & \beta u - rxu_x - zuu_y - \max_{c \geq 0} [-cu_x + U(c)] - \mathcal{L}u(z)] = 0 \\ & \forall (x, y, z) \in (0, +\infty) \times (0, +\infty) \times Z \end{aligned}$$

with

$$u(0, 0, z) = 0, \quad \forall z \in Z,$$

in the class of bounded and uniformly continuous functions.

We continue with a discussion about the history of the model.

Transaction costs are an essential feature of some economic theories, and many times are incorporated in the two-asset portfolio selection model. In [3] Constantinides assumes that the transaction costs deplete only the riskless asset and that the stock price is a logarithmic Brownian motion. He shows that if an optimal policy exists, it has to be *simple*. An investment policy is defined as simple if it is characterized by two reflecting barriers $\bar{\lambda}, \bar{\lambda}$ with $\bar{\lambda} \leq \bar{\lambda}$, such that the investor does not trade as long as the ratio y_t/x_t lies in $[\bar{\lambda}, \bar{\lambda}]$, and transacts to the closest boundary of the region of no transactions $[\bar{\lambda}, \bar{\lambda}]$ whenever this ratio lies outside this interval. He also shows that proportional transaction costs have only a second-order effect on equilibrium asset returns: the investors accommodate large transaction costs by drastically reducing the frequency and the volume of trade. Finally, he proves that the investor's expected utility of consumption is insensitive to deviations of the asset proportions from those proportions that are optimal in the absence of transaction costs. In a discrete-time version of the model, Constantinides [2], [3] proves that an optimal investment policy exists and it is simple.

In the continuous time framework, Taksar, Klass, and Assaf [16] assume that the investor does not consume but maximizes the long term expected rate of growth of wealth. In the same framework, but with more general assumptions, Fleming et al. [6] study the finite horizon problem, the average cost per unit time problem, and the growth problem and their relation.

Davis and Norman [5] relax the assumption that the transaction costs are charged only to the nonrisky asset. They consider a particular class of utility functions of the form $U(c) = c^p$ ($0 < p < 1$), and they prove that the optimal strategy confines the investor's portfolio to a certain wedge-shaped region in the portfolio plane.

Finally, there are several directions in which the two-asset problem with transaction costs can be extended. First, more than one risky asset can be allowed. Although this extension is straightforward, the computational requirements are enormous. Second, fixed transaction costs can be introduced. Some single-period models with fixed transaction costs are discussed in [1], [8], [11]–[14]. In multiperiod extensions of these models the optimal investment policy is complex, because the derived value function $u(x, y)$ is no longer homogeneous in x and y . Kandel and Ross [10] introduce quasifixed transaction costs. They use some aspects of fixed transaction costs and prove the homogeneity of the derived value function.

2. We examine some of the properties of the value function. Throughout the paper we assume

$$(2.1) \quad \beta > 2K + 1.$$

PROPOSITION 2.1. *For each $z \in Z$, u is jointly concave in x and y , strictly increasing in x , and increasing in y .*

Proof. Consider two points (x_1, y_1, z) , (x_2, y_2, z) . Let $\varepsilon > 0$ and $(C_1^\varepsilon, M_1^\varepsilon, N_1^\varepsilon)$, $(C_2^\varepsilon, M_2^\varepsilon, N_2^\varepsilon)$ be ε -optimal controls for these points respectively. Then

$$u(x_1, y_1, z) \leq E \int_0^{+\infty} e^{-\beta t} U(C_1^\varepsilon) dt + \varepsilon$$

and

$$u(x_2, y_2, z) \leq E \int_0^{+\infty} e^{-\beta t} U(C_2^\varepsilon) dt + \varepsilon.$$

Moreover, the policy $(\alpha C_1^\varepsilon + (1-\alpha)C_2^\varepsilon, \alpha M_1^\varepsilon + (1-\alpha)M_2^\varepsilon, \alpha N_1^\varepsilon + (1-\alpha)N_2^\varepsilon)$ is admissible for the point $(\alpha x_1 + (1-\alpha)x_2, \alpha y_1 + (1-\alpha)y_2, z)$. Therefore $u(\alpha x_1 + (1-\alpha)x_2, \alpha y_1 + (1-\alpha)y_2, z) \leq E \int_0^{+\infty} e^{-\beta t} U(\alpha C_1^\varepsilon + (1-\alpha)C_2^\varepsilon) dt$. Using the concavity of U , the inequalities above and sending $\varepsilon \rightarrow 0$ we conclude.

We now show that $u(\cdot, \cdot, z)$ is increasing. Consider the points (x_1, y_1, z) and (x_2, y_2, z) with $x_1 \leq x_2$, $y_1 \leq y_2$. Let $\varepsilon > 0$ and $(C^\varepsilon, M^\varepsilon, N^\varepsilon)$ be an ε -optimal policy for (x_1, y_1, z) . Since the policy $(C^\varepsilon, M^\varepsilon, N^\varepsilon)$ is admissible for the point (x_2, y_2, z) , we have

$$u(x_1, y_1, z) \leq u(x_2, y_2, z) + \varepsilon.$$

Sending $\varepsilon \rightarrow 0$ yields that $u(x_1, y_1, z) \leq u(x_2, y_2, z)$.

Finally, we show that $u(\cdot, y, z)$ is strictly increasing. To this end, let us suppose that there exist two points (x_1, y, z) and (x_2, y, z) such that $x_1 < x_2$ and $u(x_1, y, z) = u(x_2, y, z)$. Then $u(x, y, z) = u(x_1, y, z)$, for all $x \in [x_1, x_2]$. Since u is concave and nondecreasing, the interval $[x_1, x_2]$ cannot be finite. Therefore there exists a point $x_0 \geq 0$ such that $u(x, y, z) = u(x_0, y, z)$, for all $x \geq x_0$. Let $(C^\varepsilon, M^\varepsilon, N^\varepsilon)$ be an ε -optimal policy for (x_0, y, z) . Then

$$u(x_0, y, z) \leq E \int_0^{+\infty} e^{-\beta t} U(C_t^\varepsilon) dt + \varepsilon.$$

However, if $x_1 > \max(x_0, (U^{-1}[\beta(E \int_0^{+\infty} e^{-\beta t} U(C_t^\varepsilon) dt + \varepsilon)]/r)$, the policy $(rx_1, 0, 0)$ is admissible for (x_1, y, z) . Therefore

$$u(x_0, y, z) < \frac{1}{\beta} U(rx_1) = E \int_0^{+\infty} e^{-\beta t} U(rx_1) dt \leq u(x_1, y, z),$$

which contradicts our assumption. \square

PROPOSITION 2.2. *The value function u is uniformly continuous on $\bar{\Omega} = \{(x, y): x \geq 0, y \geq 0\}$.*

Proof. We first show that u is continuous on $\bar{\Omega}$. The value function is continuous in Ω , because it is concave. As a matter of fact, u is Lipschitz continuous in Ω with Lipschitz constant of order $\beta^{-1} \|U\|_\infty |(x, y)|^{-1}$.

We next show that u is continuous on the boundary. We start with the point $(0, 0)$. Since $u(0, 0, z) = 0$ (this is an immediate consequence of the assumptions in the model), we argue by contradiction.

Let us assume that for some fixed $z_0 \in Z$ there exists a positive constant M such that $\lim_{(x,y) \rightarrow (0,0)} u(x, y, z_0) = M$. Then there exists a sequence $(x_n, y_n) \rightarrow 0$ such that $u(x_n, y_n, z_0) > M/2$, for all $n \in N$. If (C^n, M^n, N^n) is an ε -optimal policy for the point (x_n, y_n, z_0) and (x_t^n, y_t^n) is the corresponding trajectory, let $w_t^n = x_t^n + (1-\mu)y_t^n$ and $w^n = x_n + (1-\mu)y_n$. Since the process M_t^n is nondecreasing we get

$$dw_t^n \leq (r + K)w_t^n dt - C_t^n dt$$

and

$$E \int_0^t e^{-\beta s} C_s^n ds \leq w^n - E[e^{\beta t} w_t^n] \leq w^n,$$

where we used (2.1) and $w_t^n \geq 0$, almost everywhere for all $t \geq 0$. From Jensen's inequality we have

$$u(x_n, y_n, z_0) - \varepsilon \leq E \int_0^{+\infty} e^{-\beta t} U(C_t^n) dt \leq \frac{1}{\beta} U(\beta w^n).$$

Sending $n \rightarrow \infty$ and using that $U(0) = 0$, we get

$$0 < \frac{M}{2} < \varepsilon.$$

Sending $\varepsilon \rightarrow 0$ we get a contradiction.

We now show that $\lim_{(x,y) \rightarrow (x_0,0)} u(x, y, z) = u(x_0, 0, z)$, for all $z \in Z$. As a matter of fact, it will be an immediate consequence of the proof that $\lim_{y \rightarrow 0} u(x, y, z) = u(x, 0, z)$ uniformly with respect to x . Consider a point $(x_0, 0, z)$ with $x_0 > 0$ fixed and a sequence (x^n, y^n, z) such that $x^n, y^n > 0$ and $\lim_{n \rightarrow \infty} (x^n, y^n) = (x_0, 0)$. Since u is locally Lipschitz it suffices to show that $\lim_{n \rightarrow \infty} |u(x_0, y^n, z) - u(x_0, 0, z)| = 0$. Finally, since u is increasing, we only need to show that $u(x_0, y^n, z) \leq u(x_0, 0, z) + \varepsilon$ for any $\varepsilon > 0$ and n sufficiently large.

Let (C^n, M^n, N^n) be an ε -optimal policy at (x_0, y^n, z) . Then

$$u(x_0, y^n, z) \leq E \int_0^{+\infty} e^{-\beta t} U(C_t^n) dt + \varepsilon.$$

Moreover, the control (C^n, \bar{M}^n, N^n) , where $d\bar{M}_t^n = dM_t^n + y^n \delta_0(t)$, is admissible for $(x_0 + (1 + \lambda)y^n, 0, z)$. Therefore

$$E \int_0^{+\infty} e^{-\beta t} U(C_t^n) dt \leq u(x_0 + (1 + \lambda)y^n, 0, z).$$

Combining the last two inequalities, we conclude. Note that all the above arguments were uniform with respect to x_0 .

We next show that $\lim_{(x,y) \rightarrow (0,y_0)} u(x, y, z) = u(0, y_0, z)$, $\forall z \in Z$. Moreover, it will be an immediate consequence of the proof that $\lim_{(x,y) \rightarrow (0,y_0)} u(x, y, z) = u(0, y_0, z)$, uniformly with respect to y .

Let $(0, y_0, z)$ with $y_0 > 0$ fixed. Arguing as before, we simply have to show that if $\varepsilon > 0$ and $x^n \rightarrow 0$ then $u(x^n, y_0, z) \leq u(0, y_0, z) + \varepsilon$.

Let (C^n, M^n, N^n) be an ε -optimal policy for (x^n, y_0, z) . Then

$$u(x^n, y_0, z) \leq E \int_0^{+\infty} e^{-\beta t} U(C_t^n) dt + \varepsilon.$$

Moreover, the policy (C^n, M^n, \bar{N}^n) is admissible for the point $(0, y_0 + (x^n/1 - \mu), z)$, where \bar{N}^n is given by $d\bar{N}_t^n = dN_t^n + (x^n/1 - \mu) \delta_0(t)$. Therefore

$$E \int_0^{+\infty} e^{-\beta t} U(C_t^n) dt \leq u\left(0, y_0 + \frac{x^n}{1 - \mu}, z\right).$$

Combining the last two inequalities, we conclude.

We now show that u is uniformly continuous on $\bar{\Omega}$.

We argue by contradiction. If u is not uniformly continuous, then there exist sequences (X_n) and (\bar{X}_n) , $X_n, \bar{X}_n \in \bar{\Omega}$, such that, as $n \rightarrow \infty$, $|X_n - \bar{X}_n| \rightarrow 0$ and

$$(2.2) \quad |u(X_n, z_0) - u(\bar{X}_n, z_0)| \geq \varepsilon$$

for some $\varepsilon > 0$ and $z_0 \in Z$.

In view of the first part of the proof, u is uniformly continuous on compact sets. Hence either (X_n) or (\bar{X}_n) , and therefore by assumption both must be unbounded. Let $X_n = (x_n, y_n)$ and $\bar{X}_n = (\bar{x}_n, \bar{y}_n)$. If $\lim_{n \rightarrow \infty} x_n > 0$ and $\lim_{n \rightarrow \infty} y_n > 0$, then the same holds for (\bar{x}_n, \bar{y}_n) . Since u is concave, locally Lipschitz with Lipschitz constant of order $|X|^{-1}$, (2.2) cannot hold.

We finally need to check what happens when either $\lim_{n \rightarrow \infty} x_n \rightarrow 0$ or $\lim_{n \rightarrow \infty} y_n = 0$. Here we only study the first case, since the other is similar. Without any loss of generality, we may assume that $\lim_{n \rightarrow \infty} x_n = 0$ and $\lim_{n \rightarrow \infty} y_n = +\infty$, otherwise we work along an appropriate subsequence. Then $\lim_{n \rightarrow \infty} \bar{x}_n = 0$ and $\lim_{n \rightarrow \infty} \bar{y}_n = +\infty$. On the other hand,

$$\begin{aligned} |u(x_n, y_n) - u(\bar{x}_n, \bar{y}_n)| &\leq |u(x_n, y_n) - u(x_n, \bar{y}_n)| + |u(x_n, \bar{y}_n) - u(\bar{x}_n, \bar{y}_n)| \\ &\leq |u(x_n, y_n) - u(x_n, \bar{y}_n)| + |u(x_n, \bar{y}_n) - u(0, \bar{y}_n)| \\ &\quad + |u(0, \bar{y}_n) - u(\bar{x}_n, \bar{y}_n)|. \end{aligned}$$

Letting $n \rightarrow \infty$ above and using the fact that u is Lipschitz continuous with respect to y uniformly with respect to x (the Lipschitz constant being of order y^{-1}) and that $\lim_{x \rightarrow \infty} u(x, y) = u(0, y)$ uniformly with respect to y , we conclude. \square

We now consider a similar control problem in which the controls, which represent the rates of trading, are assumed to be absolutely continuous processes. More precisely, we fix a positive constant L and we consider a market which offers a bond and a stock with prices evolving according to (1.1) and (1.2), respectively. The state variables x_t and y_t , which are the amount of money invested in bond and stock, obey the state equations

$$\begin{aligned} dx_t &= (rx_t - C_t) dt - (1 + \lambda)m_t dt + (1 - \mu)n_t dt, \\ dy_t &= z(t)y_t dt + m_t dt - n_t dt, \\ x_0 &= x, y_0 = y, z(0) = z. \end{aligned} \quad (2.3)$$

The controls of the investor are the *consumption rate* C_t and the *rates of trading* m_t and n_t , which are assumed to be almost everywhere bounded by L . The set of admissible controls A_L consists of controls (C, m, n) such that

- (i) C_t is F_t -measurable where $F_t = \sigma(z_s : 0 \leq s \leq t)$, $C_t \geq 0$ almost everywhere for all $t \geq 0$ and $E \int_0^{+\infty} e^{-rs} C_s ds < +\infty$.
- (ii) m_t, n_t are F_t -measurable right continuous and nonnegative processes.
- (iii) $0 \leq m_t, n_t \leq L$ almost everywhere $t \geq 0$.
- (iv) $x_t \geq 0, y_t \geq 0$ almost everywhere $t \geq 0$, where x_t, y_t are the solutions of (2.3) using the controls (C, m, n) .

The assumption that $E \int_0^{+\infty} e^{-rs} C_s ds < \infty$ is redundant here. Indeed, one can easily show that it follows from (iii) and (iv).

The control objective is to maximize the expected discounted utility from consumption over the set of admissible controls. For each fixed $L > 0$, the value function is

given by

$$u^L(x, y, z) = \sup_{A_L} E \int_0^{+\infty} e^{-\beta t} U(C_t) dt,$$

where U is the usual utility function and $\beta > 0$ is the discount factor.

PROPOSITION 2.3. *The value function u^L is jointly concave in x and y , strictly increasing in x , and increasing in y .*

Proof. The proof follows along the lines of Proposition 2.1. \square

PROPOSITION 2.4. *The value function u^L is uniformly continuous on $\bar{\Omega}$ uniformly in L .*

Proof. u^L is concave and therefore locally Lipschitz in Ω . Moreover, the Lipschitz constant is independent of L , since u^L is uniformly bounded by $\|U\|_\infty/\beta$. Therefore u^L is continuous in Ω uniformly in L . Working as in Proposition 2.2, we can prove that u^L is continuous at the point $(0, 0)$ independently of L .

We now show that u^L is continuous in $\Omega_1 = \{(x, y): x > 0, y = 0\}$. We argue by contradiction. Since u^L is locally Lipschitz in Ω_1 and nondecreasing, it suffices to assume that there exist $z_0 \in Z$ and $x_0 > 0$ such that $u^L(x_0, 0, z_0) < \lim_{y_n \rightarrow 0} u^L(x_0, y_n, z_0)$. This is equivalent to assuming that there exist $\theta > 0$ and $N_0 > 0$ such that

$$u^L(x_0, 0, z_0) + \theta \leq u^L(x_0, y_n, z_0), \quad \forall n \geq N_0.$$

On the other hand, the principle of dynamic programming gives

$$u^L(x_0, 0, z_0) \geq E \left[\int_0^\tau e^{-\beta s} U(C_s) ds + e^{-\beta \tau} u^L(x_\tau, y_\tau, z_\tau) \right]$$

for any random time τ .

Let $C_t = 0$, $m_t = 1$ and $n_t = 0$, for all $t \geq 0$, $t_n > 0$ and $\tau_n = t_n \wedge \tau_1$, where τ_1 is the first jump time of the process z_t . Then (2.3) gives

$$x_{\tau_n} = x_0 e^{r\tau_n} - \frac{1+\lambda}{r} (\exp(r\tau_n) - 1)$$

and

$$y_{\tau_n} = \frac{\exp(z_0 \tau_n) - 1}{z_0}.$$

Therefore

$$u^L(x_0, 0, z_0) \geq E \left[\exp(-\beta \tau_n) u^L \left(x_0 \exp(r\tau_n) - \frac{1+\lambda}{r} (\exp(r\tau_n) - 1), \frac{\exp(z_0 \tau_n) - 1}{z_0}, z_{\tau_n} \right) \right],$$

and, since u^L is nondecreasing

$$u^L(x_0, 0, z_0) \geq \exp(-\beta t_n) u^L \left(x_0 - \frac{1+\lambda}{r} (\exp(r t_n) - 1), \frac{\exp(z_0 t_n) - 1}{z_0}, z_0 \right) P(z_{\tau_n} = z_0).$$

We now choose t_n such that $(\exp(z_0 t_n) - 1)/z_0 = y_n$. Using that u^L is Lipschitz continuous in Ω (with the Lipschitz constant $k = k(x_0)$ independent of L), we obtain that

$$u^L(x_0, 0, z_0) \geq \exp(-\beta t_n) P(z_{\tau_n} = z_0) \left[u^L(x_0, y_n, z_0) - k \frac{1+\lambda}{r} (\exp(r t_n) - 1) \right].$$

Combining the above yields

$$u^L(x_0, 0, z_0) \geq \exp(-\beta t_n) P(z_{\tau_n} = z_0) \left[u^L(x_0, 0, z_0) + \theta - k \frac{1+\lambda}{r} (\exp(rt_n) - 1) \right].$$

We now send $y_n \rightarrow 0$. Then $\lim_{n \rightarrow \infty} t_n = 0$ and $\lim_{n \rightarrow \infty} P(z_{\tau_n} = z_0) = 1$ and $\theta \leq 0$, which is a contradiction. Therefore u^L is continuous in Ω_1 . It is also easily seen that the continuity was proved uniformly in L . Working similarly we can show that u^L is continuous in $\Omega_2 = \{(x, y): x = 0, y > 0\}$ uniformly in L . The proof that u^L is uniformly continuous on $\bar{\Omega}$ is similar to the one of Proposition 2.2 and therefore we omit it. \square

3. In this section we characterize the value functions u^L and u . We show that u^L is the unique viscosity solution of the corresponding Hamilton–Jacobi equation. We also show that the limit of u^L as $L \rightarrow \infty$ coincides with u , which is a unique viscosity solution of a system of variational inequalities. We first give the definition of viscosity solution, which was introduced by Crandall and Lions [4].

We consider a nonlinear partial differential equation of the form

$$(3.1) \quad F(X, z, u(X, z), Du(X, z)) = 0$$

where $z \in Z$, $X = (x, y)$ with $(x, y) \in \bar{\Omega}$, $Du(X, z) = (\partial u(X, z)/\partial x, \partial u(X, z)/\partial y)$ and $F: \bar{\Omega} \times Z \times \mathbb{R} \times \mathbb{R}^2 \rightarrow \mathbb{R}$ is continuous, for each $z \in Z$.

DEFINITION 3.1. A continuous function $u: \bar{\Omega} \times Z \rightarrow \mathbb{R}$ is a *constrained viscosity solution* of (3.1) if

(i) u is a *viscosity subsolution* of (3.1) on $\bar{\Omega}$, i.e., if for each $z \in Z$

$$(3.2) \quad F(X, z, u(X, z), r) \leq 0, \quad \forall X = (x, y) \in \bar{\Omega} \text{ and } r \in D_{(x,y)}^+ u(X, z),$$

where

$$D_{(x,y)}^+ u(X, z) = \left\{ r \in \mathbb{R}^2: \limsup_{h \rightarrow 0} \frac{u(X+h, z) - u(X, z) - r \cdot h}{|h|} \leq 0 \right\};$$

(ii) u is a *viscosity supersolution* of (3.1) in Ω , i.e., if for each $z \in Z$

$$(3.3) \quad F(X, z, u(X, z), r) \geq 0, \quad \forall X = (x, y) \in \Omega \text{ and } r \in D_{(x,y)}^- u(X, z),$$

where

$$D_{(x,y)}^- u(X, z) = \left\{ r \in \mathbb{R}^2: \liminf_{h \rightarrow 0} \frac{u(X+h, z) - u(X, z) - r \cdot h}{|h|} \geq 0 \right\}.$$

We now give an equivalent definition.

LEMMA 3.1. The above definition is equivalent to the following:

A continuous function $u: \bar{\Omega} \times Z \rightarrow \mathbb{R}$ is a *constrained viscosity solution* of (3.1) if

(i) u is a *viscosity subsolution* of (3.1) on $\bar{\Omega}$, i.e., if for all $\phi \in C^1(\bar{\Omega})$ at any local maximum point $X_0 \in \bar{\Omega}$ of $u - \phi$ the following holds:

$$F(X_0, z, u(X_0, z), D\phi(X_0, z)) \leq 0,$$

for each $z \in Z$.

(ii) u is a *viscosity supersolution* of (3.1) in Ω , i.e., if for all $\phi \in C^1(\Omega)$ at any local minimum point $X_0 \in \Omega$ of $u - \phi$ the following holds:

$$F(X_0, z, u(X_0, z), D\phi(X_0, z)) \geq 0$$

for each $z \in Z$.

For the proof see [4]. \square

We next show that u^L is the unique constrained viscosity solution of

$$(3.4) \quad \begin{aligned} & \beta u^L = rxu_x^L + zyu_y^L + \max_{c \geq 0} [-cu_x^L + U(c)] + \mathcal{L}u^L(z) \\ & + \max_{0 \leq m \leq L} [-(1+\lambda)u_x^L + u_y^L]m + \max_{0 \leq n \leq L} [(1-\mu)u_x^L - u_y^L]n, \\ & u^L(0, 0, z) = 0, \quad \forall z \in Z. \end{aligned}$$

This fact follows along the results of Fleming, Sethi, and Soner [7] and Soner [15], appropriately modified to deal with the generator of the process.

THEOREM 3.1. (i) u^L is a viscosity subsolution of (3.4) on $\bar{\Omega} \times Z$.

(ii) u^L is a viscosity supersolution of (3.4) in $\Omega \times Z$.

Proof. We first approximate u^L by a sequence of functions $\{u^{L,N}\}$ defined by

$$u^{L,N} = \sup_{A_{L,N}} E \int_0^{+\infty} e^{-\beta t} U(C_t) dt,$$

where

$$A_{L,N} = \{(C, m, n) \in A_L : 0 \leq C_t \leq N \text{ a.e. } \forall t \geq 0\}.$$

Working exactly as in Propositions 2.3 and 2.4, we can prove that $u^{L,N}$ is concave in (x, y) and continuous on $\bar{\Omega} \times Z$. The corresponding Hamilton–Jacobi equation is

$$(3.5) \quad \begin{aligned} & \beta u^{L,N} = rxu_x^{L,N} + zyu_y^{L,N} + \max_{N \geq c \geq 0} [-cu_x^{L,N} + U(c)] + \mathcal{L}u^{L,N}(z) \\ & + \max_{0 \leq m \leq L} [-(1+\lambda)u_x^{L,N} + u_y^{L,N}]m + \max_{0 \leq n \leq L} [(1-\mu)u_x^{L,N} - u_y^{L,N}]n. \end{aligned}$$

We first prove that $u^{L,N}$ is a viscosity subsolution of (3.5) on $\bar{\Omega}$. We will need the following lemma.

LEMMA 3.2. Let $v \in C_b(\bar{\mathcal{O}})$ be concave, where \mathcal{O} is an open subset of \mathbb{R}^n . Then

(i) $D^+v(X) \neq \emptyset, \forall X \in \mathcal{O}$

and

(ii) if $p \in D^+v(X_0)$ and $\lambda(X - X_0) + X_0 \in \bar{\mathcal{O}}, \forall X \in \bar{\mathcal{O}}$ and $\lambda \in [0, 1]$, then $v(X) \leq v(X_0) + p(X - X_0)$.

Proof. (i) Let $X_0 \in \mathcal{O}$ be fixed and consider the functions $v^\varepsilon = v * \rho_\varepsilon$, where $\varepsilon > 0$, ρ_ε is a standard mollifier and $*$ denotes convolution. Since $v^\varepsilon \rightarrow v$ as $\varepsilon \rightarrow 0$ in $B(\bar{X}_0, r) \subset \mathcal{O}$, for some $r > 0$, the functions v^ε are bounded in $B(X_0, r)$ uniformly in ε . Since the v^ε 's are also concave (recall that v is concave), the v^ε 's are also Lipschitz continuous in $B(X_0, r)$ and the Lipschitz constant is dependent of ε . By Taylor's theorem and concavity, we get

$$v^\varepsilon(X) \leq v^\varepsilon(X_0) + Dv^\varepsilon(X_0)(X - X_0), \quad \forall X \in B(X_0, r).$$

Since $|Dv^\varepsilon(X_0)| \leq C$, along subsequences $\varepsilon_n \rightarrow 0$ we have $Dv_{\varepsilon_n}(X_0) \rightarrow p$ with $p \in \mathbb{R}^N$. Letting $\varepsilon_n \rightarrow 0$ above, we get

$$v(X) \leq v(X_0) + p(X - X_0), \quad \forall X \in B(X_0, r),$$

which in turn yields that $p \in D^+v(X_0)$.

(ii) Let $p \in D^+v(X_0)$. Then

$$v(X) \leq v(X_0) + p(X - X_0) + o(|X - X_0|), \quad \forall X \in \bar{\mathcal{O}}.$$

Fix $X \in \bar{\mathcal{O}}$. Since $\lambda(X - X_0) + X_0 \in \bar{\mathcal{O}}$, for all $\lambda \in [0, 1]$, the concavity of v yields

$$v(\lambda(X - X_0) + X_0) \geq \lambda v(X) + (1 - \lambda)v(X_0).$$

Combining the last two inequalities, we get

$$\lambda v(X) + (1 - \lambda)v(X_0) \leq v(X_0) + \lambda p(X - X_0) + o(|\lambda(X - X_0)|).$$

Therefore

$$\lambda v(X) \leq \lambda v(X_0) + \lambda p(X - X_0) + o(|\lambda(X - X_0)|).$$

Dividing first by λ and then sending $\lambda \rightarrow 0$, we conclude. \square

We continue now with the proof of Theorem 3.1.

Proof. (i) In view of the definition of the constrained viscosity solution, we need to show that, if $(x_0, y_0, z) \in \bar{\Omega} \times Z$ is such that $D^+ u^{L,N}(x_0, y_0, z) \neq \emptyset$, then

$$\begin{aligned} \beta u^{L,N}(x_0, y_0, z) &\leq rx_0 p_z + zy_0 q_z + \max_{c \geq 0} [-cp_z + U(c)] + \max_{0 \leq m \leq L} [-(1 + \lambda)p_z + q_z]m \\ &\quad + \max_{0 \leq n \leq L} [(1 - \mu)p_z - q_z]n + \mathcal{L}u^{L,N}(x_0, y_0, z) \end{aligned}$$

for every $(p_z, q_z) \in D^+ u^{L,N}(x_0, y_0, z)$. To this end, assume that (x_0, y_0, z) and (p_z, q_z) are such that $(p_z, q_z) \in D^+ u^{L,N}(x_0, y_0, z)$, $z \in Z$, and define $\Phi: \mathfrak{H} \times \mathfrak{H} \times Z \rightarrow \mathfrak{H}$ by

$$\Phi(x, y, z) = u^{L,N}(x_0, y_0, z) + p_z(x - x_0) + q_z(y - y_0).$$

Lemma 3.2 yields

$$u^{L,N}(x, y, z) \leq \Phi(x, y, z), \text{ for all } (x, y, z) \in \bar{\Omega} \times Z.$$

On the other hand, the dynamic programming principle implies that for any stopping time $\tau > 0$,

$$u^{L,N}(x_0, y_0, z_0) = \sup_{A_{L,N}} E \left[\int_0^\tau e^{-\beta s} U(C_s) ds + e^{-\beta \tau} u^{L,N}(x_\tau, y_\tau, z(\tau)) \right].$$

Since $u^{L,N} \leq \Phi$, the above equality yields

$$\Phi(x_0, y_0, z_0) \leq \sup_{A_{L,N}} E \left[\int_0^\tau e^{-\beta s} U(C_s) ds + e^{-\beta \tau} \Phi(x_\tau, y_\tau, z(\tau)) \right].$$

Let θ be a positive constant and τ_1 be the first jump time of the process $z(t)$. Using Dynkin's formula and the fact that $\Phi_x(x_0, y_0, z_0) = p_{z_0}$ and $\Phi_y(x_0, y_0, z_0) = q_{z_0}$ we obtain

$$\begin{aligned} &E[e^{-\beta(\theta \wedge \tau_1)} \Phi(x_{\theta \wedge \tau_1}, y_{\theta \wedge \tau_1}, z(\theta \wedge \tau_1)) - \Phi(x_0, y_0, z_0)] \\ &= E \left[\int_0^{\theta \wedge \tau_1} e^{-\beta s} (-\beta \Phi(x_s, y_s, z_0) + rp_{z_0}x_s + z_0 q_{z_0}y_s + \mathcal{L}\Phi(x_s, y_s, z_0) \right. \\ &\quad \left. - p_{z_0}C_s + [-(1 + \lambda)p_{z_0} + q_{z_0}]m_s + [(1 - \mu)p_{z_0} - q_{z_0}]n_s) ds \right]. \end{aligned}$$

Let $\theta = 1/\ell$ and $(C^\ell, m^\ell, n^\ell) \in A_{L,N}$ be an $1/\ell^2$ -optimal policy. Then, combining the above inequalities, we get

$$\begin{aligned} -\frac{1}{\ell^2} &\leq E \int_0^{(1/\ell) \wedge \tau_1} e^{-\beta s} [U(C_s^\ell) + rp_{z_0}x_s + z_0 q_{z_0}y_s + \mathcal{L}\Phi(x_s, y_s, z_0) - p_{z_0}C_s^\ell \\ &\quad - \beta \Phi(x_s, y_s, z_0) + [-(1 + \lambda)p_{z_0} + q_{z_0}]m_s^\ell + [(1 - \mu)p_{z_0} - q_{z_0}]n_s^\ell] ds. \end{aligned}$$

On the other hand, the state equations together with the constraint $0 \leq m_t, n_t \leq L$ for almost every $t \geq 0$ give

$$|x_t^\ell - x_0| \leq (e^{rt} - 1) \left[x_0 + \frac{1 - \mu}{r} L \right],$$

$$|y_t^\ell - y_0| \leq (e^{Kt} - 1) \left[y_0 + \frac{L}{K} \right].$$

Using the above and the form of Φ , we can find a constant C_1 such that

$$\begin{aligned} -\frac{1}{\ell^2} &\leq E \int_0^{(1/\ell) \wedge \tau_1} e^{-\beta s} [rp_{z_0}x_0 + z_0q_{z_0}y_0 + \mathcal{L}\Phi(x_0, y_0, z_0) - \beta\Phi(x_0, y_0, z_0)] ds \\ &+ E \int_0^{(1/\ell) \wedge \tau_1} e^{-\beta s} [U(C_s^\ell) - p_{z_0}C_s^\ell + (-(1+\lambda)p_{z_0} + q_{z_0})m_s^\ell + ((1-\mu)p_{z_0} - q_{z_0})n_s^\ell] ds \\ &+ C_1 E \int_0^{(1/\ell) \wedge \tau_1} e^{-\beta s} \left[(e^{rs} - 1) \left(x_0 + \frac{1-\mu}{r} L \right) + (e^{Ks} - 1) \left(y_0 + \frac{L}{K} \right) \right] ds. \end{aligned}$$

Taking into account that the controls and the utility function are bounded and that $\Phi(x_0, y_0, z_0) \leq \|U\|_\infty / \beta$, we can also find a constant C_2 such that

$$\begin{aligned} -\frac{1}{\ell^2} &\leq C_2 E \int_0^{(1/\ell) \wedge \tau_1} (1 - e^{-\beta s}) ds \\ &+ C_1 E \int_0^{(1/\ell) \wedge \tau_1} e^{-\beta s} \left[(e^{rs} - 1) \left(x_0 + \frac{1-\mu}{r} L \right) + (e^{Ks} - 1) \left(y_0 + \frac{L}{K} \right) \right] ds \\ &+ E \int_0^{(1/\ell) \wedge \tau_1} [rp_{z_0}x_0 + z_0q_{z_0}y_0 + \mathcal{L}\Phi(x_0, y_0, z_0) - \beta\Phi(x_0, y_0, z_0)] ds \\ &+ E \int_0^{(1/\ell) \wedge \tau_1} U(C_s^\ell) ds - E \int_0^{(1/\ell) \wedge \tau_1} p_{z_0}C_s^\ell ds \\ &+ E \int_0^{(1/\ell) \wedge \tau_1} (-(1+\lambda)p_{z_0} + q_{z_0})m_s^\ell ds + E \int_0^{(1/\ell) \wedge \tau_1} ((1-\mu)p_{z_0} - q_{z_0})n_s^\ell ds. \end{aligned}$$

We now divide both sides by $E[(1/\ell) \wedge \tau_1]$ and we pass to the limit as $\ell \rightarrow \infty$. The first two terms will go to zero. Let

$$\begin{aligned} A_1^\ell &= \frac{1}{E((1/\ell) \wedge \tau_1)} E \int_0^{(1/\ell) \wedge \tau_1} U(C_s^\ell) ds \\ A_2^\ell &= -\frac{1}{E((1/\ell) \wedge \tau_1)} E \int_0^{(1/\ell) \wedge \tau_1} C_s^\ell ds \\ A_3^\ell &= \frac{1}{E((1/\ell) \wedge \tau_1)} E \int_0^{(1/\ell) \wedge \tau_1} (-(1+\lambda)p_{z_0} + q_{z_0})m_s^\ell ds \\ A_4^\ell &= \frac{1}{E((1/\ell) \wedge \tau_1)} E \int_0^{(1/\ell) \wedge \tau_1} ((1-\mu)p_{z_0} - q_{z_0})n_s^\ell ds. \end{aligned}$$

Let $\Gamma_{L,N} = \{U(C), C, m, n\}$ for $0 \leq C \leq N, 0 \leq m, n \leq L$. Then $(A_1^\ell, A_2^\ell, A_3^\ell, A_4^\ell) \in \overline{co}\Gamma_N$. Since the latter is a compact set, there is an element (A_1, A_2, A_3, A_4) to which $(A_1^\ell, A_2^\ell, A_3^\ell, A_4^\ell)$ converges along a subsequence. We conclude easily that

$$\begin{aligned} \beta\Phi(x_0, y_0, z_0) &\leq rx_0p_{z_0} + z_0y_0q_{z_0} + \max_{c \geq 0} [-cp_{z_0} + U(c)] + \max_{0 \leq m \leq L} [-(1+\lambda)p_{z_0} + q_{z_0}]m \\ &+ \max_{0 \leq n \leq L} [(1-\mu)p_{z_0} - q_{z_0}]n + \mathcal{L}\Phi(x_0, y_0, z_0). \end{aligned}$$

Using that $\Phi = u^{L,N}$ at (x_0, y_0, z_0) , we get (3.2).

We will next show that $\lim_{N \rightarrow \infty} u^{L,N} = u^L$ uniformly on compact subsets of $\bar{\Omega} \times Z$. We will need the following lemma.

LEMMA 3.3. Let $(C, m, n) \in A_L$ and $N > 0$. Then $(C \wedge N, m, n) \in A_{L,N}$ and

$$\lim_{N \rightarrow \infty} E \int_0^\infty e^{-\beta s} U(C_s \wedge N) ds = E \int_0^\infty e^{-\beta s} U(C_s) ds.$$

We first prove the above claim and then we present the proof of Lemma 3.3. To this end, fix $(x_0, y_0, z_0) \in \bar{\Omega} \times Z$ and let $(C^\varepsilon, m^\varepsilon, n^\varepsilon) \in A_L$ be an ε -optimal policy. Then

$$u^L(x_0, y_0, z_0) \leq E \int_0^{+\infty} e^{-\beta s} U(C_s^\varepsilon) ds + \varepsilon.$$

On the other hand, Lemma 3.3 yields that, for each $N > 0$, $(C^\varepsilon \wedge N, m^\varepsilon, n^\varepsilon) \in A_{L,N}$ and

$$E \int_0^{+\infty} e^{-\beta s} U(C_s^\varepsilon) ds \leq E \int_0^{+\infty} e^{-\beta s} U(C_s^\varepsilon \wedge N) ds + \varepsilon \text{ for } N \geq N(\varepsilon).$$

Combining the last two inequalities yields

$$u^L(x_0, y_0, z_0) \leq E \int_0^{+\infty} e^{-\beta s} U(C_s^\varepsilon \wedge N) ds + 2\varepsilon \leq u^{L,N}(x_0, y_0, z_0) + 2\varepsilon \text{ for } N \geq N(\varepsilon);$$

hence $u^{L,N}(x_0, y_0, z_0)$ is a nondecreasing sequence that converges to $u^L(x_0, y_0, z_0)$. Since both $u^{L,N}$ and u^L are continuous functions, Dini's theorem implies that $u^{L,N} \rightarrow u^L$ locally uniformly.

Proof of Lemma 3.3. The fact that $(C \wedge N, m, n) \in A_{L,N}$ follows immediately from the definitions of A_L and $A_{L,N}$.

On the other hand, since U is increasing, bounded and $U(0) = 0$,

$$0 \leq E \int_0^\infty e^{-\beta s} U(C_s) ds - E \int_0^\infty e^{-\beta s} U(C_s \wedge N) ds \leq \|U\|_\infty E \left[\int_{\{C_s \geq N\}} e^{-\beta s} ds \right].$$

To conclude we need to show that

$$\lim_{N \rightarrow \infty} E \left[\int_{\{C_s \geq N\}} e^{-\beta s} ds \right] = 0.$$

However,

$$NE \left[\int_{\{C_s \geq N\}} e^{-\beta s} ds \right] \leq E \left[\int_{\{C_s \geq N\}} e^{-\beta s} C_s ds \right] \leq E \int_0^\infty e^{-\beta s} C_s ds < \infty,$$

where the last inequality follows from the facts that $(C, m, n) \in A_L$ and $\beta > r$. Hence

$$E \left[\int_{\{C_s \geq N\}} e^{-\beta s} ds \right] \leq \frac{1}{N} E \int_0^\infty e^{-\beta s} C_s ds.$$

Letting $N \rightarrow \infty$, we conclude. \square

Finally, we show that u^L is a viscosity subsolution of (3.4) on $\bar{\Omega}$. Let $(p_{z_0}, q_{z_0}) \in D_{(x,y)}^+ u^L(x_0, y_0, z_0)$. Then there exists (cf. [4]) a smooth function $\phi: \mathfrak{R} \times \mathfrak{R} \times Z \rightarrow \mathfrak{R}$ such that $\phi_x(x_0, y_0, z_0) = p_{z_0}$, $\phi_y(x_0, y_0, z_0) = q_{z_0}$, and $u^L - \phi$ has a strict local maximum at (x_0, y_0, z_0) . Then $(u^{L,N} - \phi)(\cdot, \cdot, z_0)$ has a local maximum at (x_N, y_N, z_0) and $(x_N, y_N, z_0) \rightarrow (x_0, y_0, z_0)$. Moreover,

$$\begin{aligned} \beta u^{L,N}(x_N, y_N, z_0) &\leq rx_N p_{z_0} + z_0 y_N q_{z_0} + \max_{c \geq 0} [-cp_{z_0} + U(c)] + \max_{0 \leq m \leq L} [-(1+\lambda)p_{z_0} + q_{z_0}]m \\ &\quad + \max_{0 \leq n \leq L} [(1-\mu)p_{z_0} - q_{z_0}]n + \mathcal{L}u^{L,N}(x_N, y_N, z_0). \end{aligned}$$

Sending $N \rightarrow \infty$, we get that u^L is a viscosity subsolution of (3.4).

(ii) We now prove that $u^{L,N}$ is a supersolution of (3.4) in Ω . Let $(x_0, y_0, z_0) \in \Omega$ and $(p_{z_0}, q_{z_0}) \in D_{(x,y)}^- u^{L,N}(x_0, y_0, z_0)$. Since, by Lemma 3.2(i), $D^+ u^{L,N}(x_0, y_0, z) \neq \emptyset$,

the properties of D^+ and D^- yield that $u^{L,N}$ is differentiable in the X -direction at the point (x_0, y_0, z_0) (cf. [4]). Let Φ be defined as in (i). There exists a continuous function h with $h(0) = 0$ such that

$$u^{L,N}(x, y, z_0) \geq \Phi(x, y, z_0) + |X - X_0| h(|X - X_0|).$$

Using again the dynamic programming principle, we get

$$(3.6) \quad \Phi(x_0, y_0, z_0) \geq \sup_{A_{L,N}} E \left[\int_0^\theta e^{-\beta s} U(C_s) ds + e^{-\beta\theta} \Phi(x_\theta, y_\theta, z_\theta) + e^{-\beta\theta} |X_\theta - X_0| h(|X_\theta - x_0|) \right].$$

Let $(C, m, n) \in A_{L,N}$ with $C_t = C_0$, $m_t = m_0$, $n_t = n_0$, $\forall t \geq 0$ and $\tau = \theta \wedge \tau_1$ where $\theta = T \wedge \inf\{\tau > 0: x_\tau^0 = 0\} \wedge \inf\{\tau > 0: y_\tau^0 = 0\}$ and x_τ^0, y_τ^0 are the corresponding trajectories. Using Dynkin's formula and (3.6), we get

$$\begin{aligned} 0 &\geq C_1 E \int_0^{\theta \wedge \tau_1} e^{-\beta s} \left[(e^{rs} - 1) \left(x_0 + \frac{1-\mu}{r} L \right) + (e^{Ks} - 1) \left(y_0 + \frac{L}{K} \right) \right] ds \\ &\quad + C_2 E \int_0^{\theta \wedge \tau_1} (1 - e^{-\beta s}) ds \\ &\quad + E \int_0^{\theta \wedge \tau_1} [U(C_0) - p_{z_0} C_0 + r x_0 p_{z_0} + z_0 y_0 q_{z_0} + [-(1+\lambda)p_{z_0} + q_{z_0}] m_0 \\ &\quad + [(1-\mu)p_{z_0} - q_{z_0}] n_0 + \mathcal{L}\Phi(x_0, y_0, z_0) - \beta\Phi(x_0, y_0, z_0)] ds \end{aligned}$$

for some constants C_1 and C_2 . Dividing by $E[\theta \wedge \tau_1]$, sending $T \rightarrow 0$ and using that $u^{L,N}(x_0, y_0, z_0) = \Phi(x_0, y_0, z_0)$ we obtain (3.3). Finally, working similarly as in (i) we can show that u^L is a viscosity supersolution of (3.4) in $\bar{\Omega}$. \square

THEOREM 3.2. *Let u and v be bounded, uniformly continuous such that u is a viscosity subsolution of (3.4) on $\bar{\Omega}$ and v is a viscosity supersolution of (3.4) in Ω . Then $u \leq v$ on $\bar{\Omega}$.*

Proof. Let $X = (x, y) \in \bar{\Omega}$, $P = (p, q) \in R \times R$ and $H: \bar{\Omega} \times Z \times R^2 \rightarrow R$ be given by

$$H(X, z, P) = rxp + zyq + F(p) + \max_{0 \leq m \leq L} [(1+\lambda)p - q]m + \max_{0 \leq n \leq L} [-(1-\mu)p + q]n,$$

where $F(p) = \max_{c \geq 0} [-cp + U(c)]$, $p > 0$.

We argue by contradiction; i.e., we assume that

$$(3.7) \quad \max_{z \in Z} \sup_{X \in \bar{\Omega}} [u(X, z) - v(X, z)] > 0.$$

Then for sufficiently small $\theta > 0$

$$(3.8) \quad \max_{z \in Z} \sup_{X \in \bar{\Omega}} [u(X, z) - v(X, z) - \theta|X|^2] > 0.$$

Indeed, if not, there would be a sequence $\theta_n \downarrow 0$ such that $\max_{z \in Z} \sup_{X \in \bar{\Omega}} [u(X, z) - v(X, z) - \theta_n|X|^2] \leq 0$, which in turn yields $\max_{z \in Z} \sup_{X \in \bar{\Omega}} [u(X, z) - v(X, z)] \leq 0$, contradicting (3.7).

Since the process z takes a finite number of values and u and v are bounded, we can find points $z_0 \in Z$ and $\bar{X} \in \bar{\Omega}$ such that

$$(3.9) \quad u(\bar{X}, z_0) - v(\bar{X}, z_0) - \theta|\bar{X}|^2 = \max_{z \in Z} \sup_{X \in \bar{\Omega}} [u(X, z) - v(X, z) - \theta|X|^2].$$

In what follows we omit z_0 .

Next, for $\varepsilon > 0$ we define $\psi: \bar{\Omega} \times \bar{\Omega} \rightarrow \mathbb{R}$ by

$$\psi(X, Y) = u(X) - v(Y) - \left| \frac{Y - X}{\varepsilon} - 4(1, 1) \right|^2 - \theta|X|^2,$$

and we claim that ψ attains its maximum at a point, say (X_0, Y_0) , such that for ε small and some $\ell > 0$

$$(3.10) \quad |Y_0 - X_0| \leq \ell\varepsilon.$$

Indeed, observe that ψ is bounded. Let (X_n, Y_n) be a maximizing sequence. Then

$$\lim_{n \rightarrow \infty} \left[u(X_n) - v(Y_n) - \left| \frac{Y_n - X_n}{\varepsilon} - 4(1, 1) \right|^2 - \theta|X_n|^2 \right] = \sup_{\bar{\Omega} \times \bar{\Omega}} \psi(X, Y) < +\infty.$$

However,

$$(3.11) \quad \sup_{\bar{\Omega} \times \bar{\Omega}} \psi(X, Y) \geq \psi(\bar{X}, \bar{X} + 4\varepsilon(1, 1)) \geq u(\bar{X}) - v(\bar{X}) - \theta|\bar{X}|^2 - \omega_v(k\varepsilon),$$

where ω_v is the modulus of continuity of v and $k > 0$. Using (3.8) and (3.9), we see that for ε sufficiently small

$$(3.12) \quad \sup_{\bar{\Omega} \times \bar{\Omega}} \psi(X, Y) > 0.$$

We also observe that if

$$u(X_n) - v(Y_n) - \left| \frac{Y_n - X_n}{\varepsilon} - 4(1, 1) \right|^2 \leq 0$$

as $n \rightarrow \infty$ we contradict (3.12). Therefore

$$u(X_n) - v(Y_n) \geq \left| \frac{Y_n - X_n}{\varepsilon} - 4(1, 1) \right|^2,$$

which implies (3.10). On the other hand, the choice of (X_n, Y_n) and (3.12) yield that the sequence (X_n) , and, in view of the above observation, (Y_n) are bounded as $n \rightarrow \infty$. Hence, along subsequences, (X_n, Y_n) converge to a maximum point of ψ , which we denote by (X_0, Y_0) .

Moreover, (3.10) and (3.11) give

$$(3.13) \quad \left| \frac{Y_0 - X_0}{\varepsilon} - 4(1, 1) \right|^2 \leq \omega_n(k\varepsilon) + \omega_v(\ell\varepsilon).$$

We can choose ε small such that $\omega_n(k\varepsilon) + \omega_v(\ell\varepsilon) \leq 1$. Then there is a vector e with $|e| \leq 1$ such that $Y_0 = X_0 + 4\varepsilon(1, 1) + \varepsilon e$, which implies that $Y_0 \in \Omega$.

We now consider the functions

$$\begin{aligned} \phi(Y) &= u(X_0) - \left| \frac{Y - X_0}{\varepsilon} - 4(1, 1) \right|^2 - \theta|X_0|^2, \\ \bar{\phi}(X) &= v(Y_0) + \left| \frac{Y_0 - X}{\varepsilon} - 4(1, 1) \right|^2 + \theta|X|^2. \end{aligned}$$

Since $u - \bar{\phi}$ has a maximum at $X_0 \in \bar{\Omega}$ and $v - \phi$ has a minimum at $Y_0 \in \Omega$, applying the definition of viscosity solution as in Lemma 3.1, we get

$$\beta u(X_0, z_0) \leq H(X_0, z_0, P_\varepsilon + 2\theta X_0) + \sum_{z_0 \neq z'} q_{z_0 z'} [u(X_0, z) - u(X_0, z_0)]$$

and

$$\beta v(Y_0, z_0) \geq H(Y_0, z_0, P_\varepsilon) + \sum_{z_0 \neq z'} q_{z_0 z'} [v(Y_0, z) - v(Y_0, z_0)],$$

where

$$P_\varepsilon = -\frac{2}{\varepsilon} \left(\frac{Y_0 - X_0}{\varepsilon} - 4(1, 1) \right).$$

Combining the above inequalities yields

$$\begin{aligned} \beta[u(X_0, z_0) - v(Y_0, z_0)] &\leq [H(X_0, z_0, P_\varepsilon + 2\theta X_0) - H(Y_0, z_0, P_\varepsilon)] \\ (3.14) \quad &+ \sum_{z_0 \neq z'} q_{z_0 z'} [u(X_0, z) - u(X_0, z_0) - v(Y_0, z) + v(Y_0, z_0)]. \end{aligned}$$

On the other hand, from the definition of H , we have that

$$\begin{aligned} (3.15) \quad &|H(X_0, z_0, P_\varepsilon + 2\theta X_0) - H(Y_0, z_0, P_\varepsilon)| \\ &\leq |H(X_0, z_0, P_\varepsilon + 2\theta X_0) - H(X_0, z_0, P_\varepsilon)| + K|X_0 - Y_0||P_\varepsilon| \end{aligned}$$

for some $K > 0$. Let $X_0 = (x_0, y_0)$ and $P_\varepsilon = (p_\varepsilon, q_\varepsilon)$. Then

$$\begin{aligned} &H(x_0, y_0, z_0, 2\theta x_0 + p_\varepsilon, 2\theta y_0 + q_\varepsilon) - H(x_0, y_0, z_0, p_\varepsilon, q_\varepsilon) \leq 2\theta[rx_0^2 + zy_0^2] \\ &+ \left[\max_{0 \leq m \leq L} [(1+\lambda)(2\theta x_0 + p_\varepsilon) - (2\theta y_0 + q_\varepsilon)]m - [(1+\lambda)p_\varepsilon - q_\varepsilon]L \right] \\ &+ \left[\max_{0 \leq n \leq L} [-(1-\mu)(2\theta x_0 + p_\varepsilon) + (2\theta y_0 + q_\varepsilon)]n - [-(1-\mu)p_\varepsilon + q_\varepsilon]L \right] \\ &\leq C\theta L|X_0| + 2\theta[rx_0^2 + zy_0^2] \leq \theta|X_0|^2 + \theta C^2 L^2 + 2\theta[rx_0^2 + zy_0^2] \\ &\leq \beta\theta|X_0|^2 + C^2 L^2 \theta, \end{aligned}$$

for some $C > 0$, where we used that $K \geq r$, (2.1), and that F is a decreasing function. Using the above inequality and (3.10), (3.13), and (3.15), we get

$$\begin{aligned} (3.16) \quad &|H(X_0, z_0, P_\varepsilon + 2\theta X_0) - H(Y_0, z_0, P_\varepsilon)| \\ &\leq 2K\ell[\omega_v(\ell\varepsilon) + \omega_v(k\varepsilon)]^{1/2} + \beta\theta|X_0|^2 + C^2 L^2 \theta. \end{aligned}$$

Moreover, from (3.10) we have

$$u(X_0, z) - v(X_0, z) \leq u(X_0, z_0) - v(Y_0, z_0) + \omega_v(k\varepsilon),$$

which combined with (3.11) gives

$$[u(X_0, z) - u(X_0, z_0)] - [v(Y_0, z) - v(Y_0, z_0)] \leq \omega_v(k\varepsilon) + \omega_v(\ell\varepsilon).$$

Finally, using that $0 \leq q_{zz'} \leq 1$ and $\sum_{z' \in Z} q_{zz'} = 1$, we get

$$(3.17) \quad \mathcal{L}u(X_0, z_0) - \mathcal{L}v(Y_0, z_0) \leq \omega_v(k\varepsilon) + \omega_v(\ell\varepsilon).$$

Using now (3.14), (3.16), and (3.17), we have

$$\begin{aligned} \beta[u(X_0, z_0) - v(Y_0, z_0) - \theta|X_0|^2] &\leq 2K\ell[\omega_v(\ell\varepsilon) + \omega_v(k\varepsilon)]^{1/2} \\ &+ C^2 L^2 \theta + \omega_v(k\varepsilon) + \omega_v(\ell\varepsilon), \end{aligned}$$

and, using the definition of (X_0, Y_0) ,

$$\max_{\Omega \times \Omega} \psi(X, Y) \leq \frac{1}{\beta} \left[2K\ell[\omega_v(\ell\varepsilon) + \omega_v(k\varepsilon)]^{1/2} + \omega_v(k\varepsilon) + \omega_v(\ell\varepsilon) \right] + \frac{C^2 L^2 \theta}{\beta}.$$

Then, however, (3.9) and (3.11) yield

$$\begin{aligned} & \max_{z \in Z} \sup_{X \in \bar{\Omega}} [u(X, z) - v(X, z) - \theta |X|^2] \\ & \leq \omega_v(k\varepsilon) + \frac{1}{\beta} [2K\ell[\omega_v(\ell\varepsilon) + \omega_v(k\varepsilon)]^{1/2} + \omega_v(k\varepsilon) + \omega_v(\ell\varepsilon)] + \frac{C^2 L^2}{\beta} \theta, \end{aligned}$$

which, in turn, implies that

$$u(X, z) - v(X, z) - \theta |X|^2 \leq \frac{C^2 L^2}{\beta} \theta \quad \forall X \in \bar{\Omega} \text{ and } \forall z \in Z.$$

Letting $\theta \rightarrow 0$ contradicts (3.7). \square

Remark 3.1. Although it was not necessary for the above proof, but it will be used later on, we show that

$$(3.18) \quad \lim_{\theta \downarrow 0} \lim_{\varepsilon \downarrow 0} \theta |X_0(\theta, \varepsilon)|^2 = 0.$$

Indeed, from (3.11) we have

$$\psi(X_0, X_0) + \omega_v(k\varepsilon) \geq u(\bar{X}) - v(\bar{X}) - \theta |\bar{X}|^2 - \omega_v(k\varepsilon),$$

which yields

$$(3.19) \quad u(X_0, z_0) - v(X_0, z_0) - \theta |X_0|^2 \geq [u(\bar{X}, z_0) - v(\bar{X}, z_0) - \theta |\bar{X}|^2] - 2\omega_v(k\varepsilon),$$

and, in turn $\sup_{\varepsilon > 0} |X_0(\theta, \varepsilon)| < \infty$.

Therefore there exists $\hat{X}_0(\theta)$ such that $\lim_{\varepsilon \rightarrow 0} |X_0(\theta, \varepsilon)|^2 = \hat{X}_0(\theta)$, otherwise we contradict (3.19). The limit here is taken along subsequences, which to simplify notation we denote the same way as the whole family. By sending $\varepsilon \downarrow 0$, (3.19) combined with (3.9) implies

$$(3.20) \quad u(\hat{X}_0, z_0) - v(\hat{X}_0, z_0) - \theta |\hat{X}_0|^2 \geq u(X, z) - v(X, z) - \theta |X|^2 \quad \forall X \in \bar{\Omega}, \forall z \in Z.$$

We now send $\theta \rightarrow 0$. If $\lim_{\theta \downarrow 0} \theta |\hat{X}_0|^2 = \alpha \neq 0$, again along subsequences, (3.20) yields

$$\max_{z \in Z} \sup_{\bar{\Omega}} [u(X, z) - v(X, z)] - \alpha \geq \max_{z \in Z} \sup_{\bar{\Omega}} [u(X, z) - v(X, z)].$$

Therefore $\max_{z \in Z} \sup_{\bar{\Omega}} [u(X, z) - v(X, z)] < 0$, which contradicts (3.7).

PROPOSITION 3.1. As $L \rightarrow \infty$, $u^L \rightarrow w \in C(\bar{\Omega})$.

Proof. Fix $(x_0, y_0) \in \bar{\Omega}$. The sequence $u^L(x_0, y_0)$ is increasing as $L \rightarrow \infty$, therefore $\lim_{L \rightarrow \infty} u^L(x_0, y_0) = w(x_0, y_0)$ exists. Moreover, since the functions u^L are continuous at (x_0, y_0) uniformly in L , w is continuous on $\bar{\Omega}$. \square

4. In this section we prove that w coincides with the value function $u \in C(\bar{\Omega})$. We first show that u is a constrained viscosity solution of a certain variational inequality and second that this variational inequality has a unique constrained viscosity solution.

THEOREM 4.1. The value function u is a constrained viscosity solution of

$$(4.1) \quad \begin{aligned} & \min [(1 + \lambda)u_x - u_y, -(1 - \mu)u_x + u_y, \beta u - rxu_x - zy u_y - F(u_x) - \mathcal{L}u(z)] = 0 \\ & \forall (x, y, z) \in (0, +\infty) \times (0, +\infty) \times Z \text{ with } u(0, 0, z) = 0, \forall z \in Z. \end{aligned}$$

Proof. We first show that u is viscosity subsolution of (4.1) on $\bar{\Omega}$. To this end, let (x_0, y_0, z_0) be fixed with $(x_0, y_0) \in \bar{\Omega}$, consider $(p_{z_0}, q_{z_0}) \in D_{x,y}^+ u(x_0, y_0, z_0)$ and define $\Phi: \Re \times \Re \times Z \rightarrow \Re$ by

$$\Phi(x, y, z) = u(x_0, y_0, z) + p_z(x - x_0) + q_z(y - y_0),$$

where $(p_z, q_z) \in D_{(x,y)}^+ u(x_0, y_0, z)$. Lemma 3.2 yields

$$(4.2) \quad u(x, y, z) \leq \Phi(x, y, z).$$

We are going to show that

$$(4.3) \quad \min [(1+\lambda)p_{z_0} - q_{z_0}, -(1-\mu)p_{z_0} + q_{z_0}, \beta u - rx_0p_{z_0} - z_0y_0q_{z_0} - F(p_{z_0}) - \mathcal{L}u(z_0)] \leq 0$$

where $F(p) = \max_{c \geq 0} [-cp + U(c)]$.

If $(1+\lambda)p_{z_0} - q_{z_0} \leq 0$ or $-(1-\mu)p_{z_0} + q_{z_0} \leq 0$, the above inequality is obvious. So let us assume that

$$(4.4) \quad (1+\lambda)p_{z_0} - q_{z_0} > 0 \text{ and } -(1-\mu)p_{z_0} + q_{z_0} > 0.$$

In the following, we are first going to assume that the control C is such that $0 \leq C_t \leq N$ for almost every $t \geq 0$, and then we will remove the upper bound. Since the arguments are similar to the ones used in Theorem 3.1, we proceed as if there is no bound on C . Later, we will mention when we use this upper bound.

Applying the dynamic programming principle at the point (x_0, y_0, z_0) with stopping time $\theta = (1/\ell) \wedge \tau_1 = \min\{1/\ell, \tau_1\}$, where τ_1 is the first jump time of z_t , we obtain

$$(4.5) \quad u(x_0, y_0, z_0) \leq E \left[\int_0^\theta e^{-\beta s} U(C_s^\ell) ds + e^{-\beta\theta} u \left(e^{r\theta} \left(x_0 - \int_0^\theta e^{-rs} C_s^\ell ds - (1+\lambda)m_\theta^\ell + (1-\mu)n_\theta^\ell \right), e^{z_0\theta} (y_0 + \hat{m}_\theta^\ell - \hat{n}_\theta^\ell), z_\theta \right) \right] + \frac{1}{\ell^2},$$

where (C^ℓ, M^ℓ, N^ℓ) is an $1/\ell^2$ -optimal policy and

$$m_\theta^\ell = \int_0^\theta e^{-rs} dM_s^\ell, \quad \hat{m}_\theta^\ell = \int_0^\theta \exp(-z_0 s) dM_s^\ell$$

and

$$n_\theta^\ell = \int_0^\theta e^{-rs} dN_s^\ell, \quad \hat{n}_\theta^\ell = \int_0^\theta \exp(-z_0 s) dN_s^\ell.$$

Let $r \geq z_0$ (the case $r < z_0$ is treated similarly). Then

$$(4.6) \quad m_\theta^\ell \leq \hat{m}_\theta^\ell \text{ and } n_\theta^\ell \leq \hat{n}_\theta^\ell.$$

Since the control (C^ℓ, M^ℓ, N^ℓ) is admissible, we also have

$$(4.7) \quad x_0 - \int_0^\theta e^{-rs} C_s^\ell ds \geq (1+\lambda)m_\theta^\ell - (1-\mu)n_\theta^\ell$$

and

$$(4.8) \quad y_0 \geq -\hat{m}_\theta^\ell + \hat{n}_\theta^\ell.$$

Moreover,

$$(4.9) \quad \hat{m}_\theta^\ell \leq (\exp(r - z_0)\theta) \int_0^\theta e^{-rs} dM_s^\ell \leq \left(\exp\left(\frac{r - z_0}{\ell}\right) \right) m_\theta^\ell$$

and, similarly,

$$(4.10) \quad \hat{n}_\theta^\ell \leq \left(\exp\left(\frac{r - z_0}{\ell}\right) \right) n_\theta^\ell.$$

From (4.7)–(4.10) we get

$$(4.11) \quad m_\theta^\ell \leq \frac{x_0 + (1-\mu)y_0}{c},$$

where $c > 0$ is such that $(1 + \lambda) - (1 - \mu) \exp((r - z_0)/\ell) \geq c$ for ℓ sufficiently large. Similarly,

$$(4.12) \quad n_\theta^\ell \leq \frac{\exp((r - z_0)/\ell)x_0 + (1 + \lambda)y_0}{c} \leq \frac{2x_0 + (1 + \lambda)y_0}{c}$$

for ℓ sufficiently large.

Moreover, since

$$m_\theta^\ell = \int_0^\theta e^{-rs} dM_s^\ell \geq \exp(-r/\ell)(M_\theta^\ell - M_0^\ell),$$

using (4.11) we obtain

$$(4.13) \quad M_\theta^\ell - M_0^\ell \leq \exp(r/\ell) \frac{x_0 + (1 - \mu)y_0}{c} \leq k_1$$

for some $k_1 > 0$, and, similarly

$$(4.14) \quad N_\theta^\ell - N_0^\ell \leq k_2$$

for some $k_2 > 0$.

Therefore there exist constants \bar{C} , K_1 and K_2 such that

$$(4.15) \quad \hat{m}_\theta^\ell - m_\theta^\ell \leq E \int_0^\theta \bar{C}s dM_s^\ell \leq \frac{K_1}{\ell}$$

and

$$(4.16) \quad \hat{n}_\theta^\ell - n_\theta^\ell \leq \frac{K_2}{\ell}.$$

Using that u is a nondecreasing function, $r \geq z_0$, (4.5), and (4.15), we have

$$(4.17) \quad \begin{aligned} u(x_0, y_0, z_0) &\leq \|U\|_\infty E(\theta) + E[e^{-\beta\theta} u(e^{r\theta}(x_0 + \bar{X}_\theta), e^{r\theta}(y_0 + \bar{Y}_\theta), z_\theta) \\ &\quad - e^{-\beta\theta} u(x_0 + \bar{X}_\theta, y_0 + \bar{Y}_\theta, z_0)] \\ &\quad + E[e^{-\beta\theta} u(x_0 + \bar{X}_\theta, y_0 + \bar{Y}_\theta, z_0)], \end{aligned}$$

where

$$\bar{X}_\theta = -(1 + \lambda)\hat{m}_\theta^\ell + \frac{(1 + \lambda)K_1}{\ell} + (1 - \mu)\hat{n}_\theta^\ell$$

and

$$\bar{Y}_\theta = \hat{m}_\theta^\ell - \hat{n}_\theta^\ell.$$

Let ω be the modulus of continuity of u . Then

$$(4.18) \quad \begin{aligned} &E[e^{-\beta\theta} u(e^{r\theta}(x_0 + \bar{X}_\theta), e^{r\theta}(y_0 + \bar{Y}_\theta), z_\theta) - e^{-\beta\theta} u(x_0 + \bar{X}_\theta, y_0 + \bar{Y}_\theta, z_0)] \\ &\leq K_3[\|u\|_\infty P(z_\theta \neq z_0) + \omega(e^{r\theta} - 1)] \end{aligned}$$

for some positive constant K_3 .

Moreover, (4.2), (4.17), and (4.18) yield

$$\begin{aligned} u(x_0, y_0, z_0) &\leq \|U\|_\infty E(\theta) + K_3[\|u\|_\infty P(z_\theta \neq z_0) + \omega(e^{r\theta} - 1)] \\ &\quad + E e^{-\beta\theta} [u(x_0, y_0, z_0) + p_{z_0} \bar{X}_\theta + q_{z_0} \bar{Y}_\theta]. \end{aligned}$$

Since $P[z_\theta \neq z_0] = 0(\theta)$, we get

$$\begin{aligned} & E[(1+\lambda)p_{z_0} - q_{z_0}]\hat{m}_\theta^\ell + (-(1-\mu)p_{z_0} + q_{z_0})\hat{n}_\theta^\ell \\ & \leq K_3[\|u\|_\infty E(\theta) + \omega(e^{r\theta} - 1)] + \|U\|_\infty E(\theta) + \frac{(1+\lambda)K_1}{\ell}. \end{aligned}$$

Finally, in view of (4.4), we can find positive constants M_1 and M_2 such that

$$E(\hat{m}_\theta^\ell) \leq M_1 \left[E(\theta) + \omega\left(\frac{1}{\ell}\right) + \frac{1}{\ell} \right]$$

and

$$E(\hat{n}_\theta^\ell) \leq M_2 \left[E(\theta) + \omega\left(\frac{1}{\ell}\right) + \frac{1}{\ell} \right].$$

From (4.2) and (4.5) we get

$$u(x_0, y_0, z_0) \leq E \int_0^\theta e^{-\beta s} U(C_s^\ell) ds + E[e^{-\beta\theta} \Phi(x_\theta^\ell, y_\theta^\ell, z_\theta)] + \frac{1}{\ell^2},$$

where $x_\theta^\ell, y_\theta^\ell$ are given by (1.3) with control (C^ℓ, M^ℓ, N^ℓ) . Using Dynkin's formula, we have

$$\begin{aligned} u(x_0, y_0, z_0) & \leq E \int_0^\theta e^{-\beta s} U(C_s^\ell) ds + u(x_0, y_0, z_0) \\ & \quad + E \int_0^\theta e^{-\beta s} [-\beta \Phi(x_s^\ell, y_s^\ell, z_0) + rx_s^\ell p_{z_0} + z_0 y_s^\ell q_{z_0} - C_s^\ell p_{z_0} \\ & \quad + \mathcal{L}\Phi(x_s^\ell, y_s^\ell, z_0)] ds + E \int_0^\theta e^{-\beta s} [-(1+\lambda)p_{z_0} + q_{z_0}] dM_s^\ell \\ & \quad + E \int_0^\theta e^{-\beta s} [(1-\mu)p_{z_0} - q_{z_0}] dN_s^\ell + \frac{1}{\ell^2}. \end{aligned}$$

Since M_t and N_t are nondecreasing processes, using (4.4), we get

$$\begin{aligned} 0 & \leq E \int_0^\theta e^{-\beta s} U(C_s^\ell) ds + E \int_0^\theta [rp_{z_0}(x_s^\ell - x_0) + z_0 q_{z_0}(y_s^\ell - y_0)] ds \\ & \quad + E \int_0^\theta e^{-\beta s} [-\beta u(x_0, y_0, z_0) + rp_{z_0} x_0 \\ & \quad + z_0 q_{z_0} y_0 - p_{z_0} C_s^\ell + \mathcal{L}\Phi(x_s^\ell, y_s^\ell, z_0)] ds + \frac{1}{\ell^2}. \end{aligned}$$

Let

$$A(\theta) = E \int_0^\theta [rp_{z_0}(x_s^\ell - x_0) + z_0 q_{z_0}(y_s^\ell - y_0)] ds$$

and

$$B(\theta) = E \int_0^\theta e^{-\beta s} [-\beta u(x_0, y_0, z_0) + rx_0 p_{z_0} + z_0 y_0 q_{z_0} - p_{z_0} C_s^\ell + \mathcal{L}\Phi(x_s^\ell, y_s^\ell, z_0)] ds.$$

Then

$$A(\theta) = E \int_0^\theta (rp_{z_0}[(e^{rs} - 1)x_0 - (1 + \lambda)e^{rs}m_s^\ell + (1 - \mu)e^{rs}n_s^\ell] \\ + z_0q_{z_0}[(\exp(z_0s) - 1)y_0 + (\exp(z_0s))\hat{m}_s^\ell - (\exp(z_0s))\hat{n}_s^\ell]) ds.$$

Since $E \int_0^\theta e^{hs}k_s^\ell ds \leq (1/\ell)e^{h/\ell}E(k_\theta^\ell) \leq (K/\ell)e^{h/\ell}[E(\theta) + \omega(1/\ell) + 1/\ell]$, where h is r or z_0 and k_s^ℓ is m_s^ℓ , \hat{m}_s^ℓ , n_s^ℓ , or \hat{n}_s^ℓ , we have

$$(4.19) \quad \lim_{\ell \rightarrow \infty} \frac{1}{E(\theta)} E \int_0^\theta e^{hs}k_s^\ell ds = 0.$$

Therefore $\lim_{\ell \rightarrow \infty} A(\theta)/E(\theta) = 0$.

Relations (4.7), (4.12), and $\beta > r$ give

$$(4.20) \quad \int_0^\theta e^{-\beta s} C_s^\ell ds \leq x_0 + \frac{(1 - \mu)[2x_0 + (1 + \lambda)y_0]}{c}.$$

On the other hand,

$$\begin{aligned} \mathcal{L}\Phi(x_s^\ell, y_s^\ell, z_0) &= \sum_{z' \neq z} q_{z'z} [\Phi(x_s^\ell, y_s^\ell, z') - \Phi(x_s^\ell, y_s^\ell, z_0)] \\ &= \sum_{z' \neq z} q_{z'z} [u(x_0, y_0, z') + p_{z'}(x_s^\ell - x_0) + q_{z'}(y_s^\ell - y_0) \\ &\quad - u(x_0, y_0, z_0) - p_{z_0}(x_s^\ell - x_0) - q_{z_0}(y_s^\ell - y_0)]. \end{aligned}$$

Using that Z is a finite set and (4.19), we get

$$(4.21) \quad \lim_{\ell \rightarrow \infty} \frac{E \int_0^\theta e^{-\beta s} \mathcal{L}\Phi(x_s^\ell, y_s^\ell, z_0) ds}{E(\theta)} = \mathcal{L}u(x_0, y_0, z_0).$$

Finally, from (4.19)–(4.21) we conclude that

$$\beta u(x_0, y_0, z_0) \leq rx_0p_{z_0} + z_0y_0q_{z_0} + \mathcal{L}u(x_0, y_0, z_0) + F(p_{z_0}).$$

This last conclusion follows along the lines of the analogous argument in the proof of Theorem 3.1. To complete the proof, we need to remove the bound on C_t , which can be done again as in Theorem 3.1.

We finally show that u is a supersolution of (4.1) in Ω . To this end, fix $(x_0, y_0, z_0) \in \Omega \times Z$ and consider $(p_{z_0}, q_{z_0}) \in D_{(x,y)}^- u(x_0, y_0, z_0)$. Then there exists a smooth function $\psi: \mathbb{R} \times \mathbb{R} \times Z \rightarrow \mathbb{R}$ such that $u - \psi$ has a strict minimum at (x_0, y_0, z_0) , $u(x_0, y_0, z_0) = \psi(x_0, y_0, z_0)$ and $p_{z_0} = \psi_x(x_0, y_0, z_0)$, $q_{z_0} = \psi_y(x_0, y_0, z_0)$. Then

$$(4.22) \quad \psi(x_0, y_0, z_0) \geq \sup_A E \left[\int_0^\theta e^{-\beta s} U(C_s) ds + e^{-\beta\theta} \psi(x_\theta, y_\theta, z_\theta) \right].$$

In particular, if we use a control (C, M, N) such that $C_t = 0$, for all $t \geq 0$, $M_0 = M$, $N_0 = N$ and $M_t = N_t = 0$, for all $t > 0$ in (4.22), we get

$$\psi(x_0, y_0, z_0) \geq \psi(x_0 - (1 + \lambda)M + (1 - \mu)N, y_0 + M - N, z_0).$$

This yields

$$(4.23) \quad \min [(1 + \lambda)\psi_x(x_0, y_0, z_0) - \psi_y(x_0, y_0, z_0), -(1 - \mu)\psi_x(x_0, y_0, z_0) \\ + \psi_y(x_0, y_0, z_0)] \geq 0.$$

Now if we use a constant control $(C_0, 0, 0)$ in (4.22), we obtain

$$\psi(x_0, y_0, z_0) \geq E \left[\int_0^\theta e^{-\beta s} U(C_s) ds + e^{-\beta \theta} \psi(x_\theta, y_\theta, z_\theta) \right],$$

where $\theta = (1/\ell) \wedge \inf \{\tau: x_\tau^0 = 0\}$, where $\ell > 0$ and x_τ^0 is the corresponding trajectory. We proceed as in Theorem 3.1 and we obtain

$$(4.24) \quad \beta u(x_0, y_0, z_0) \geq rx_0 p_{z_0} + z_0 y_0 q_{z_0} + F(p_{z_0}) + \mathcal{L}u(x_0, y_0, z_0).$$

Combining (4.23) and (4.24), we get

$$\begin{aligned} \min [(1+\lambda)p_{z_0} - q_{z_0}, -(1-\mu)p_{z_0} - q_{z_0}, \beta u - rx_0 p_{z_0} - z_0 y_0 q_{z_0} \\ - F(p_{z_0}) - \mathcal{L}u(x_0, y_0, z_0)] \geq 0. \end{aligned} \quad \square$$

THEOREM 4.2. *The variational inequality (4.1) has a unique constrained viscosity solution in the class of bounded uniformly continuous functions.*

Proof. We are going to show that if v and u are respectively a supersolution in Ω and a subsolution on $\bar{\Omega}$ of (4.1) then $v \geq u$ on $\bar{\Omega}$. To end this, we follow the strategy of Ishii [9]. Let $\phi: \bar{\Omega} \rightarrow \Re$ be defined by $\phi(x, y) = C_1 x + C_2 y + k$, where C_1, C_2, k are positive constants satisfying

$$(1-\mu)C_1 < C_2 < (1+\lambda)C_1$$

and

$$\beta k > rC_1 + KC_2 + F(C_1).$$

Let $X = (x, y) \in \bar{\Omega}$, $P = (p, q) \in \Re \times \Re$ and $H: \bar{\Omega} \times Z \times \Re \times \Re^2 \rightarrow \Re$ given by

$$H(X, z, v, P) = \min [(1+\lambda)p - q, -(1-\mu)p + q, \beta v - rxp - z y q - F(p) - \mathcal{L}v].$$

An easy calculation shows that there exists a positive constant M such that

$$(4.25) \quad H(X, z, \phi, \nabla \phi) \geq M > 0, \quad \forall X \in \bar{\Omega} \setminus \{0, 0\}, z \in Z.$$

Let $\theta \in (0, 1)$ and define $v_\theta = \theta v + (1-\theta)\phi$ on $\bar{\Omega} \times Z$. The functions v_θ are bounded and uniformly continuous. Moreover, since the map $(v, P) \rightarrow H(X, z, v, P)$ is concave, ϕ satisfies (4.25) and v is a supersolution in Ω , we have that

$$(4.26) \quad \begin{aligned} H(X, z, v_\theta, \nabla v_\theta) &\geq \theta H(X, z, v, \nabla v) + (1-\theta)H(X, z, \phi, \nabla \phi) \\ &\geq M(1-\theta) > 0, \quad \forall (X, z) \in \Omega \times Z \end{aligned}$$

holds in the viscosity sense. Therefore v_θ is a supersolution of $H(X, z, v_\theta, \nabla v_\theta) = M(1-\theta)$ in Ω . We now need the following lemma.

LEMMA 4.1. *Let $w, v: \bar{\Omega} \times Z \rightarrow \Re$ be uniformly continuous and nondecreasing with respect to x . If w is a subsolution of $H(X, z, w, \nabla w) = 0$ on $\bar{\Omega}$ and v is a supersolution of $H(X, z, v, \nabla v) = c$ in Ω for some $c > 0$, v is bounded from below and w is bounded, then $v \geq w$ on $\bar{\Omega}$.*

Proof. Since the proof is similar to the one of Theorem 3.2, we only show the main steps. We assume that

$$(4.27) \quad \max_{z \in Z} \sup_{X \in \bar{\Omega}} [w(X, z) - v(X, z)] > 0.$$

This implies that for sufficiently small $\theta > 0$

$$(4.28) \quad \max_{z \in Z} \sup_{X \in \bar{\Omega}} [w(X, z) - v(X, z) - \theta |X|^2] > 0.$$

We can find points $z_0 \in Z$ and $\bar{X} \in \bar{\Omega}$ such that

$$(4.29) \quad w(\bar{X}, z_0) - v(\bar{X}, z_0) - \theta|\bar{X}|^2 = \max_{z \in Z} \sup_{X \in \bar{\Omega}} [w(X, z) - v(X, z) - \theta|X|^2].$$

Next, we consider the function $\psi: \bar{\Omega} \times \bar{\Omega} \rightarrow \mathbb{R}$ with

$$\psi(X, Y) = w(X, z_0) - v(Y, z_0) - \left| \frac{Y - X}{\varepsilon} - 4(1, 1) \right|^2 - \theta|X|^2$$

for $\varepsilon > 0$ and we look at its maximum denoted by (X_0, Y_0) . Working as in Theorem 3.2 we can show that $Y_0 \in \Omega$. Moreover, from Remark 3.1 we have that

$$(4.30) \quad \lim_{\theta \downarrow 0} \lim_{\varepsilon \downarrow 0} \theta|X_0|^2 = 0.$$

We now consider the functions

$$\begin{aligned} \phi(Y) &= w(X_0) - \left| \frac{Y - X_0}{\varepsilon} - 4(1, 1) \right|^2 - \theta|X_0|^2, \\ \bar{\phi}(X) &= v(Y_0) + \left| \frac{Y_0 - X}{\varepsilon} - 4(1, 1) \right|^2 + \theta|X|^2. \end{aligned}$$

Since $w - \bar{\phi}$ has a maximum at X_0 and $v - \phi$ has a minimum at Y_0 , from Lemma 3.1 we have

$$H(X_0, z_0, w(X_0, z_0), P_\varepsilon + 2\theta X_0) \leq 0$$

and

$$H(Y_0, z_0, v(Y_0, z_0), P_\varepsilon) \geq c,$$

where

$$P_\varepsilon = -\frac{2}{\varepsilon} \left(\frac{Y_0 - X_0}{\varepsilon} - 4(1, 1) \right).$$

Combining the above inequalities yields

$$(4.31) \quad H(X_0, z_0, w(X_0, z_0), P_\varepsilon + 2\theta X_0) - H(Y_0, z_0, v(Y_0, z_0), P_\varepsilon) \leq -c.$$

Let $X_0 = (x_0, y_0)$, $Y_0 = (\bar{x}_0, \bar{y}_0)$ and $P_\varepsilon = (p_\varepsilon, q_\varepsilon)$. Using the form of H , (4.31) becomes

$$\begin{aligned} (4.32) \quad & \min [(1 + \lambda)(p_\varepsilon + 2\theta x_0) - (q_\varepsilon + 2\theta y_0), -(1 - \mu)(p_\varepsilon + 2\theta x_0) + (q_\varepsilon + 2\theta y_0), \\ & \beta w - r x_0(p_\varepsilon + 2\theta x_0) - z_0 y_0(q_\varepsilon + 2\theta y_0) - F(p_\varepsilon + 2\theta x_0) - \mathcal{L}w(X_0, z_0)] \\ & - \min [(1 + \lambda)p_\varepsilon - q_\varepsilon, -(1 - \mu)p_\varepsilon + q_\varepsilon, \beta v - r \bar{x}_0 p_\varepsilon - z_0 \bar{y}_0 q_\varepsilon \\ & - F(p_\varepsilon) - \mathcal{L}v(Y_0, z_0)] \leq -c. \end{aligned}$$

We now look at the following cases.

Case (i). $H(X_0, z_0, w(X_0, z_0), P_\varepsilon + 2\theta X_0) = (1 + \lambda)(p_\varepsilon + 2\theta x_0) - (q_\varepsilon + 2\theta y_0)$. Then (4.32) yields

$$(1 + \lambda)2\theta x_0 - 2\theta y_0 \leq -c.$$

Using (4.30), we get a contradiction.

Case (ii). $H(X_0, z_0, x(X_0, z_0), P_\varepsilon + 2\theta X_0) = -(1 - \mu)(p_\varepsilon + 2\theta x_0) + (q_\varepsilon + 2\theta y_0)$.
In this case, (4.32) yields

$$2\theta[-(1 - \mu)x_0 + y_0] \leq -c,$$

which contradicts (4.30).

Case (iii).

$$\begin{aligned} H(X_0, z_0, w(X_0, z_0), P_\varepsilon + 2\theta X_0) &= \beta w - rx_0(p_\varepsilon + 2\theta x_0) - z_0 y_0(q_\varepsilon + 2\theta y_0) \\ &\quad - F(p_\varepsilon + 2\theta x_0) - \mathcal{L}w(X_0, z_0). \end{aligned}$$

Then (4.32) yields

$$\begin{aligned} \beta[w(X_0, z_0) - v(Y_0, z_0)] + c &\leq r(x_0 - \bar{x}_0)p_\varepsilon + z_0(y_0 - \bar{y}_0)q_\varepsilon + F(p_\varepsilon + 2\theta x_0) \\ &\quad - F(p_\varepsilon) + \mathcal{L}w(X_0, z_0) - \mathcal{L}v(Y_0, z_0). \end{aligned}$$

Working similarly as in the proof of Theorem 3.2, we get

$$\beta[w(X_0, z_0) - v(Y_0, z_0)] + c \leq \theta |X_0|^2 + 2K\ell[\omega_v(\ell\varepsilon) + \omega_v(k\varepsilon)]^{1/2} + \omega_v(k\varepsilon) + \omega_v(\ell\varepsilon).$$

Again working as in the proof of Theorem 3.2, sending first $\varepsilon \rightarrow 0$, then $\theta \rightarrow 0$, and using (4.30), we contradict (4.27). \square

Acknowledgment. This work is part of the author's Ph.D. dissertation under the direction of Professor W. H. Fleming, whom I thank for his suggestions and advice.

REFERENCES

- [1] M. J. BRENNAN, *The optimal number of securities in a risky asset portfolio when there are fixed costs of transacting: theory and some empirical results*, J. Financial Quantitative Anal., 10 (1975), pp. 483-496.
- [2] G. M. CONSTANTINIDES, *Multiperiod consumption and investment behavior with convex transactions costs*, Management Sci., 25 (1979), pp. 1127-1137.
- [3] ———, *Capital market equilibrium with transaction costs*, J. Political Econom., 94 (1986), pp. 842-862.
- [4] M. G. CRANDALL AND P.-L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1-42.
- [5] M. H. A. DAVIS AND A. R. NORMAN, *Portfolio selection with transaction costs*, Math. Oper. Res., 15 (1990), pp. 676-713.
- [6] W. H. FLEMING, S. GROSSMAN, J.-L. VILA, AND T. ZARIPHPOULOU, *Optimal portfolio rebalancing with transaction costs*, Econometrica, submitted.
- [7] W. H. FLEMING, S. SETHI, AND M. H. SONER, *An optimal stochastic planning problem with randomly fluctuating demand*, SIAM J. Control Optim., 25 (1987), pp. 1494-1502.
- [8] D. GOLDSMITH, *Transaction costs and the theory of portfolio selection*, J. Finance, 31 (1976), pp. 1127-1139.
- [9] H. ISHII, *A simple, direct proof of uniqueness for solutions of the Hamilton-Jacobi equations of Eikonal type*, Proc. Amer. Math. Soc., 100 (1987), pp. 247-251.
- [10] S. KANDEL, AND S. A. ROSS, *Some intertemporal models of portfolio selection with transaction costs*, Working Paper 107, University of Chicago, Grad. School Bus., Center Res. Security Prices, Chicago, IL, 1983.
- [11] H. E. LELAND, *On consumption and portfolio choices with transaction costs*, in Essays on Economic Behavior under Uncertainty, M. Balch, D. McFadden, and S. Wu, eds., North-Holland, Amsterdam (1974).
- [12] H. LEVY, *Equilibrium in an imperfect market: A constraint on the number of securities in the portfolio*, Amer. Econom. Rev., 68 (1978), pp. 643-658.
- [13] J. MAYSHAR, *Transaction costs in a model of capital market equilibrium*, J. Political Econom., 87 (1979), pp. 673-700.
- [14] R. MUKHERJEE AND E. ZABEL, *Consumption and portfolio choices with transaction costs*, in Essays on Economic Behavior under Uncertainty, M. Balch, D. McFadden, and S. Wu, eds., North-Holland, Amsterdam, 1974.

- [15] M. H. SONER, *Optimal control with state-space constraints*, SIAM J. Control Optim., 24 (1986), pp. 552–561.
- [16] M. TAKSAR, M. J. KLASS, AND D. ASSAF, *A diffusion model for optimal portfolio selection in the presence of brokerage fees*, Dept. of Statistics, Florida State University, Tallahassee, FL, 1986.
- [17] T. ZARIPHOPULOU, *Optimal investment-consumption models with constraints*, Ph.D. thesis, Brown University, Providence, RI, 1988.

ESTIMATION OF UNKNOWN VARIABLE PARAMETERS IN MOVING BOUNDARY PROBLEMS*

K. A. MURPHY†

Abstract. The problem of estimating unknown variable parameters appearing in moving boundary problems is considered; these are specifically nonlinear diffusion equations defined on a moving spatial domain. A spline-based approximation method that results in a sequence of computationally tractable approximate parameter estimation problems has been developed. A convergence result is proved for a certain class of these moving boundary problems. The paper is concluded with a set of representative numerical examples.

Key words. parameter estimation, moving boundary problems, spline approximation

AMS(MOS) subject classifications. 65, 69

1. Introduction. Moving boundary problems model phenomena in diverse areas. Many examples can be found in the papers collected in [7], [10], [12], and [29] (this represents just a sample). We consider explicit one-dimensional moving boundary problems; this type consists of a parabolic partial differential equation defined on a moving domain (of one spatial dimension), coupled to an ordinary differential equation describing the movement of the boundary. Within this type of moving boundary problem, we consider fairly general model equations. The coefficients we propose to estimate can be functions of time, space, or the state variables. We do not require a priori parametrizations of nonconstant parameters.

Moving boundary problems present interesting theoretical questions, as can be seen from surveying the many papers and books devoted to theoretical aspects (again, we give just a small sample of the available literature: [7], [12]–[14], [29]). Similarly, much attention has been paid to the development of numerical algorithms for the forward problem: Given fixed parameters, find a numerical approximation to the solution of the moving boundary problem (see, for example, [1], [11], [20], [22]).

Our concern here is the development of computational and theoretical ideas for the solution of the parameter estimation problem: Given observations of a process that we assume can be modelled by a moving boundary problem, we would like to estimate unknown variable parameters within the model equations. For the Stefan problem, specifically, several methods (see, e.g., [15], [17]) have been proposed for parameter estimation and control problems (both are often referred to as “inverse Stefan problems”); for a survey, see [16] and the references therein. Here, we develop both a theoretical framework and a corresponding computational algorithm, applicable to a general type of model equation (we note that the Stefan problem is not included in this theoretical framework; however, our numerical algorithm has been successfully applied to a test Stefan problem).

Our approach to the parameter estimation problem is similar in spirit to that of [2]–[6]. For a given set of model equations containing unknown parameters, we pose the problem as one of minimizing a least squares fit-to-data. As the evaluation of the

* Received by the editors February 21, 1989; accepted for publication (in revised form) March 18, 1991. This research was supported in part by Air Force Office of Scientific Research contract AFOSR-86-0256. Parts of the research were carried out while the author was a visitor at the Institute for Computer Applications in Science and Engineering (ICASE), National Aeronautics and Space Administration (NASA) Langley Research Center, Hampton, Virginia, which is operated under NASA contract NAS1-18107.

† Department of Mathematics, University of North Carolina, Chapel Hill, North Carolina 27599.

objective function requires solution of an infinite-dimensional problem (in that the states are the solution of the coupled partial and ordinary differential equations, and the parameters are, in general, functions), we replace the original problem with a sequence of approximating problems. It is then necessary to prove that the parameters we estimate with the computationally tractable, finite-dimensional approximate problems converge in some sense to best-fit parameters for the original model of interest.

This paper is structured as follows. We describe our model equations in § 2, defining there the class of problems to which our theoretical results apply. In § 3 we discuss several motivating example problems. These model equations arise from various biological and engineering applications. Three of our example problems belong to the class for which we have demonstrated both theoretical and numerical results; for two examples, we present only numerical results (although we do have some preliminary theoretical results for the Stefan problem; this is in preparation).

In § 4 we restate our equations in variational form (see, e.g., [2], [6]), and formulate our approximations. We precisely state all of our hypotheses, both on solutions of the original model equations as well as on our approximations. Our main convergence results are then stated in § 5 (with the proofs presented in the Appendix).

We turn to numerical considerations in §§ 6 and 7. The numerical implementation is described in § 6, and the results of numerical test examples are presented in § 7.

We use the following notation throughout the paper: $|\cdot|$ denotes absolute value, $|\cdot|_\infty$ designates the $L^\infty(\Omega)$ norm (where the domain Ω should be clear from the context), and $\langle \cdot, \cdot \rangle$, $\|\cdot\|$ represent the inner product and norm in $L^2(0, 1)$.

2. The parameter estimation problem. We consider general model equations of the following form:

$$\begin{aligned}
 (2.1) \quad & \frac{\partial u}{\partial t} = \frac{\partial}{\partial x} \left(D \frac{\partial u}{\partial x} + Vu \right) + \rho(u) + f(t, x), \quad 0 < x < s(t), \quad 0 < t \leq T; \\
 & \alpha_{11} \left(D \frac{\partial u}{\partial x} + V_1 u \right) \Big|_{x=0} - \alpha_{12} u(t, 0) = g(t); \\
 & \alpha_{21} \left(D \frac{\partial u}{\partial x} + V_1 u \right) \Big|_{x=s(t)} + \alpha_{22} u(t, s(t)) = h(t), \quad 0 < t \leq T; \\
 & u(0, x) = u_0(x), \quad 0 \leq x \leq s_0; \\
 & \frac{ds}{dt} = \mathcal{F}(s(t), u(t, \cdot); \gamma), \quad 0 \leq t \leq T; \\
 & s(0) = s_0.
 \end{aligned}$$

Parameter spaces will be defined more precisely below; generally, the diffusion term is assumed to be a function of both t and x . The convective term V is assumed to have two components: $V \equiv V_1(t, x) + V_2(\dot{s})$ (this is motivated by one of the applications examples to be discussed in the next section). We allow the boundary terms α_{ij} to be time dependent (in fact, we let α_{22} depend on \dot{s} in one special case), satisfying $\alpha_{ij} \geq 0$ for i and $j = 1, 2$; we require that α_{i1} ($i = 1, 2$) is either identically zero or nonzero for all $t \in [0, T]$, and that $\alpha_{i1}^2 + \alpha_{i2}^2 \neq 0$. The functional \mathcal{F} , describing the dynamics of the

moving boundary, will, in general, depend on the states of the system (s and u), as well as an additional parameter(s) designated by γ ; this parameter might be a function of time, location, and/or the states. We classify moving boundary problems according to \mathcal{F} ; we discuss this in more detail below.

There are many papers in the literature that address the existence and regularity of solutions of moving boundary problems, primarily those of Stefan type (see, e.g., [13], [14], [29]); however, there is as yet no general theory that can be applied to a class of problems. Our interest here is in the parameter estimation problem and, in particular, the question of convergence of our approximation scheme. For a given model equation, we assume that a weak solution exists, and that by appropriately restricting the parameter sets, the solution will possess sufficient smoothness for our convergence theory (this is precisely stated in § 4). We prove that weak solutions are unique, thereby ensuring that our least squares fit-to-data functional is well defined.

Suppose that we have data corresponding to measurements of u and/or s . The data could represent various quantities, depending on the experiment, but we assume that they correspond to the following type of observations:

$$\hat{u}_i \sim \int_0^{s(t_i)} u(t_i, x) dx \quad \text{and} \quad \hat{s}_i \sim s(t_i), \text{ (or } \hat{s}_i \sim s(t_i) - s_0), \quad \text{for } i = 1, \dots, m,$$

where (u, s) is a solution of the above model equations. This is the form of data to which our theory, developed in § 5, can be directly applied. Another likely form of measurement is $u(t_i, 0)$ or $u(t_i, s(t_i))$, (or, perhaps less natural, we might have a set of observations of u distributed throughout the spatial domain, i.e., $u(t_i, x_j)$). Given data of this form (i.e., discrete spatial measurements) we would need a stronger convergence statement than that obtained in § 5 to make our theory complete. We do include some numerical test examples using such forms of data, and observe that our method is still successful. This suggests that our methods may be useful for a broader set of circumstances than that covered by our theory.

In addition to the data, we assume that f and some subset of the remaining parameters are known. The method we discuss in this paper can be used to estimate any of D , V_1 , V_2 , ρ , α_{ij} , g , h , γ , u_0 , and s_0 . Clearly, in practice, we could not hope to simultaneously estimate all of the parameters listed from one set of data; for the purpose of discussion, however, and to show the full capabilities of the method, we proceed as though all parameters are unknown. We expect the data to be noisy and, moreover, the model equations cannot be considered exact. Therefore, we pose the estimation problem as one of minimizing a least squares fit-to-data criterion.

The inverse problem suffers, as do all inverse problems, from certain theoretical and numerical difficulties. Of particular worry is the lack of uniqueness of solutions and the lack of continuous dependence of the parameters on the data. Such issues, while of obvious import, are not the focus of the investigations reported here. They are discussed in [9] and [18], for example. As described in [2], the use of compactness constraints on the admissible parameter sets can mitigate these difficulties (both theoretical and numerical). By imposing compactness assumptions on our parameter sets, we are able to prove a theorem regarding convergence of our approximations, as well as ensuring that the “method stability” defined in [2] holds.

It will be convenient to have the model equations defined on a fixed spatial domain, so we make the change of variables (see, e.g., [10, p. 187]) $y = x/(s(t))$, and define the new state $U(t, y) = u(t, x(y))$, the forcing function $F(t, y) = f(t, x(y))$, and the parameters $\mathcal{D}(t, y) = D(t, x(y))$, $\mathcal{V}_1(t, y) = V_1(t, x(y))$, $\mathcal{V}_2(\dot{s}) = V_2(\dot{s})$ (no change of variable is required; however, we keep our notation consistent), $U_0(y) = u_0(x(y))$,

$\Gamma(t, y, s, U) = \gamma(t, x(y), s, u(t, x(y)))$, and $\tilde{\mathcal{F}}(s, U; \Gamma) = \mathcal{F}(s, u; \gamma)$, thereby obtaining

$$\frac{\partial U}{\partial t} = \frac{1}{s^2} \frac{\partial}{\partial y} \left(\mathcal{D} \frac{\partial U}{\partial y} \right) + \frac{1}{s} \frac{\partial}{\partial y} (\mathcal{V}_1 U) + \frac{1}{s} \mathcal{V}_2 \frac{\partial U}{\partial y} + \frac{\dot{s}}{s} y \frac{\partial U}{\partial y} + \rho(U) + F(t, y), \quad 0 < y < 1,$$

$$0 < t \leq T;$$

$$\begin{aligned} & \alpha_{11} \left(\frac{\mathcal{D}}{s} \frac{\partial U}{\partial y} + \mathcal{V}_1 U \right) \Big|_{y=0} - \alpha_{12} U(t, 0) = g(t); \\ (2.2) \quad & \alpha_{21} \left(\frac{\mathcal{D}}{s} \frac{\partial U}{\partial y} + \mathcal{V}_1 U \right) \Big|_{y=1} + \alpha_{22} U(t, 1) = h(t), \quad 0 < t \leq T; \\ & U(0, y) = U_0(y), \quad 0 \leq y \leq 1; \\ & \frac{ds}{dt} = \tilde{\mathcal{F}}(s, U; \Gamma), \quad 0 \leq t \leq T; \\ & s(0) = s_0. \end{aligned}$$

We defer the discussion of the compactness assumptions required for the parameter sets to § 4; there we define the set \mathcal{Q}_c , a compact subset of the underlying space \mathcal{X}_q . Let $q = (\mathcal{D}, \mathcal{V}_1, \mathcal{V}_2, \rho, \alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22}, g, h, \Gamma, V_0, s_0)$ designate the vector of all parameters. Let $\Omega \equiv [0, T] \times [0, 1]$ and, given any space \mathbb{X} , let $(\mathbb{X})^m$ represent the product of m copies of \mathbb{X} . While the specific form of the functional $\tilde{\mathcal{F}}: \mathbb{R} \times L^\infty(0, 1) \times \mathcal{G} \rightarrow \mathbb{R}$ and the definition of the set \mathcal{G} are problem dependent, we always assume that \mathcal{G} is a subset of some normed space \mathcal{X}_γ (we consider three examples in § 3). Given the positive constants K_s and \bar{u} , we define the space \mathcal{X}_q as

$$\mathcal{X}_q \equiv (C(\Omega))^2 \times C[-K_s, K_s] \times C[0, \bar{u}] \times (C[0, T])^6 \times \mathcal{X}_\gamma \times C(0, 1) \times \mathbb{R}.$$

We then write $\mathcal{Q} = \mathcal{D} \times \mathcal{B}_1 \times \mathcal{B}_2 \times \mathcal{P} \times \mathbb{I}_{11} \times \mathbb{I}_{12} \times \mathbb{I}_{21} \times \mathbb{I}_{22} \times \mathcal{G} \times \mathcal{S} \times \mathcal{G} \times \mathcal{I}_v \times \mathcal{I}_s$, where we assume that $\mathcal{Q} \subset \mathcal{X}_q$ is a bounded subset of \mathcal{X}_q . Given the positive constants D_L , ρ^{Lip} , and $\mathcal{V}_2^{\text{Lip}}$, we further assume that

$$\begin{aligned} \mathcal{D} & \subset \{\mathcal{D} \in C(\Omega) \mid \mathcal{D}(t, y) \geq D_L\}, \\ \mathcal{B}_2 & \subset \{\mathcal{V}_2 \in C[-K_s, K_s] \mid |\mathcal{V}_2(\theta_1) - \mathcal{V}_2(\theta_2)| \leq \mathcal{V}_2^{\text{Lip}} |\theta_1 - \theta_2|\}, \\ \mathcal{P} & \subset \{\rho \in C[0, \bar{u}] \mid |\rho(\theta_1) - \rho(\theta_2)| \leq \rho^{\text{Lip}} |\theta_1 - \theta_2|\}, \\ \mathbb{I}_{ij} & \subset \{\alpha \in C[0, T] \mid \alpha(t) \geq 0\} \quad \text{for each } i, j = 1, 2. \end{aligned}$$

We make the following general assumptions about $\tilde{\mathcal{F}}$:

(H \mathcal{F}) $\tilde{\mathcal{F}}: \mathbb{R} \times L^\infty(0, 1) \times \mathcal{G} \rightarrow \mathbb{R}$ is continuous in all arguments, and for each

$\Gamma \in \mathcal{G}$, $\tilde{\mathcal{F}}$ satisfies

$$|\tilde{\mathcal{F}}(s, U_1; \Gamma) - \tilde{\mathcal{F}}(s, U_2; \Gamma)| \leq \lambda(|s|, \Gamma) |U_1 - U_2|_\infty,$$

$$|\tilde{\mathcal{F}}(s_1, U; \Gamma) - \tilde{\mathcal{F}}(s_2, U; \Gamma)| \leq \mu(|U|_\infty, \Gamma) |s_1 - s_2|,$$

where λ and μ are continuous in both arguments.

Assumption (H \mathcal{F}) is the means by which we classify moving boundary problems. It is for such model equations that we prove a convergence theorem for approximations of the parameter estimation problem. In the next section, we give several examples that fall into this class. We note that this class excludes some well-known moving boundary problems, such as the Stefan problem and the oxygen diffusion problem

(see, e.g., [10], [21]); our numerical scheme, however, has nevertheless proved successful for these problems (see § 7).

3. Motivating examples. In this section we describe three models arising from a variety of applications areas, each of which can be seen to belong to the class of moving boundary problems defined above. We also describe two models that do not fit into our theoretical framework, but for which we have had success with preliminary numerical test examples.

3.1. Biofilm on activated carbon particles. In [8] a model is derived to describe the process by which contaminants diffuse through a liquid and bind to activated carbon particles. As described in [8] (we refer the interested reader to this paper and the references cited there for a full discussion), a biofilm layer, containing bacteria, forms in a thin layer around the carbon particles. The contaminant must therefore diffuse through this layer to attach to the carbon. The contaminant is a source of nutrient for the bacteria, and as the bacteria utilize the contaminant, the biofilm layer grows. Simultaneously, the biofilm layer may decrease in thickness due to shear loss. The full model developed in [8] consists of a coupled system of partial and ordinary differential equations describing the dynamics in the surrounding bulk liquid, the biofilm layer, and in the carbon particles.

We have simplified this model, in a way that extracts the moving boundary dynamics; this simplified version has been briefly discussed in [24]. Essentially, we uncouple the biofilm dynamics from the rest of the problem by assuming that certain boundary parameters are known functions of time. Here in our notation is the simplified model, where $u(t, x)$ represents the concentration of contaminant in the biofilm layer, $s(t)$ represents the width of the layer with $x = 0$ corresponding to the carbon-biofilm interface, and $x = s$ corresponds to the biofilm-liquid interface:

$$\begin{aligned}
 (3.1) \quad & \frac{\partial u}{\partial t} = \frac{\partial}{\partial x} \left(D \frac{\partial u}{\partial x} \right) - \rho(u), \quad t \in (0, T], \quad x \in (0, s(t)); \\
 & \left(D \frac{\partial u}{\partial x} \right) \Big|_{x=0} = g(t); \\
 & \left(D \frac{\partial u}{\partial x} \right) \Big|_{x=s(t)} + \alpha u(t, s(t)) = h(t), \quad t \in (0, T]; \\
 & u(0, x) = u_0(x), \quad x \in [0, s_0]; \\
 & \frac{ds}{dt} = \int_0^{s(t)} (\gamma_1(u(t, x)) - \gamma_2(t)) \, dx, \quad t \in [0, T]; \\
 & s(0) = s_0.
 \end{aligned}$$

In [8] the nonlinear terms ρ and γ_1 have been parametrized, with $\gamma_1(u) = Yku/(K + u)$ and $\rho = (1/Y)\gamma_1$. We can estimate these terms without any such a priori assumptions, but we do assume that they exhibit similar behavior; for example, we assume that there exist constants K_γ and γ_1^{Lip} such that γ_1 satisfies $|\gamma_1| \leq K_\gamma$ and $|\gamma_1(u_1) - \gamma_1(u_2)| \leq \gamma_1^{\text{Lip}}|u_1 - u_2|$. We also assume that $\gamma_2 \in C[0, T]$. Making the change of variables to fixed domain, we find that $\tilde{\mathcal{F}}(s, U; \Gamma) \equiv s \int_0^1 (\gamma_1(U(t, y)) - \gamma_2(t)) \, dy$. It is easily verified that $\tilde{\mathcal{F}}$ satisfies assumption (H $\tilde{\mathcal{F}}$) of the previous section, with

$$\lambda(|s|, \Gamma) = |s| \gamma_1^{\text{Lip}} \quad \text{and} \quad \mu(|U|_\infty, \Gamma) = \left(K_\gamma + \sup_{t \in [0, T]} |\gamma_2| \right).$$

For this example, given the constant \bar{u} , we let $\mathcal{X}_\gamma = C[0, \bar{u}] \times C[0, T]$ and suppose that \mathcal{G} is a bounded subset of \mathcal{X}_γ satisfying $\mathcal{G} \subset \{(\gamma_1, \gamma_2) \in \mathcal{X}_\gamma \mid |\gamma_1| \leq K_\gamma \text{ and } |\gamma_1(u_1) - \gamma_1(u_2)| \leq \gamma_1^{\text{Lip}}|u_1 - u_2|\}$.

3.2. Acrosomal elongation. We consider a model derived by Perelson and Coutsias [26], specifically (15)–(20). For a detailed description of the model, we refer the interested reader to [26]; we will describe it briefly here. In the process of fertilizing an egg, the sperm must extend a long tube called the acrosomal process. It is still not completely understood how this structure is created on the timescale necessary. Perelson and Coutsias present a moving boundary model that describes one possible mechanism. By analyzing their model, as they have presented it, they conclude the proposed mechanism does not adequately model the observed phenomenon.

Nevertheless, we feel that this moving boundary model can provide a beginning for study, with parameter estimation techniques providing a valuable tool. We let $u(t, x)$ represent the concentration of unbound monomer and $s(t)$ represent the length of the acrosomal process at any time t . The model equations are

$$\begin{aligned} \frac{\partial u}{\partial t} &= D \frac{\partial^2 u}{\partial x^2} - s \frac{\partial u}{\partial x}, & t \in (0, T], \quad x \in (0, s(t)); \\ u(t, 0) &= u_0; \\ (3.2) \quad D \frac{\partial u}{\partial x}(t, s(t)) &+ k_1 \alpha(t) u(t, s(t)) = k_{-1} \alpha(t), & t \in (0, T]; \\ u(0, x) &= u_0, & x \in [0, s_0]; \\ \frac{ds}{dt} &= \gamma_1 u(t, s(t)) - \gamma_2, & t \in [0, T]; \\ s(0) &= s_0. \end{aligned}$$

While our notation is different, we have essentially reproduced the equations of Perelson and Coutsias, except that we allow our boundary coefficient α to be time varying. While Perelson and Coutsias acknowledge that this parameter should realistically be nonconstant, it is unknown whether this change is enough to make the model realistic.

The parameters γ_1 and γ_2 are assumed to be constants. After the change of variables, we obtain $\tilde{\mathcal{F}}(s, U; \Gamma) \equiv \gamma_1 U(t, 1) - \gamma_2$, and it is trivial to verify that assumption (H $\tilde{\mathcal{F}}$) holds with $\lambda(|s|, \Gamma) = |\gamma_1|$, and $\mu(|U|_\infty, \Gamma) \equiv 0$. For this example, \mathcal{G} is assumed to be a bounded subset of $\mathcal{X}_\gamma = \mathbb{R}^2$.

3.3. Glassy polymers. This model problem comes from [13]. The equations derived there describe the morphological change caused by the presence of a solvent in a polymer, when that solvent exists at a sufficiently high concentration. For the details of the model derivation, we refer the interested reader to [13] and the references therein. In brief, the solvent diffuses through the polymer, which initially has a glassy quality. When the solvent reaches a sufficiently high concentration, the polymer instantly assumes a gel-like quality.

This model of Fasano and Ricci has been discussed in some detail in the context of parameter estimation in [25]. There, the original model equations of [13] are changed slightly (we retain the diffusion constant explicitly in the model equations and shift time by an arbitrarily small amount to obtain nonzero initial conditions). In our notation, u represents the concentration of the solvent in the glassy region, and s

represents the location of discontinuity between the two regions. Our model equations are

$$\begin{aligned}
 (3.3) \quad & \frac{\partial u}{\partial t} = D \frac{\partial^2 u}{\partial x^2}, \quad t \in (0, T], \quad x \in (0, s(t)); \\
 & u(t, 0) = 1; \\
 & D \frac{\partial u}{\partial x}(t, s(t)) + \dot{s}(t)u(t, s(t)) = -\dot{s}(t)\sigma(t), \quad t \in (0, T]; \\
 & u(0, x) = u_0(x), \quad x \in [0, s_0]; \\
 & \frac{ds}{dt} = \gamma(u(t, s(t))), \quad t \in [0, T]; \\
 & s(0) = s_0.
 \end{aligned}$$

We direct the reader's attention to the boundary condition at $x = s(t)$; this is nonlinear and requires a redefinition of the set \mathbb{U}_{12} and \mathfrak{S} , since the parameters α_{12} and h depend on \dot{s} . In fact, this nonlinearity makes the convergence arguments of § 5 delicate, requiring more than just the redefinition of parameter sets; we pursue this in § 5.

We obtain $\tilde{\mathcal{F}}(s, U; \Gamma) \equiv \gamma(U(t, 1))$ after the change of variables; if we assume that γ satisfies $\gamma \in C[0, \bar{u}]$ and $|\gamma(u_1) - \gamma(u_2)| \leq \gamma^{\text{Lip}}|u_1 - u_2|$, then it is easy to show that $\tilde{\mathcal{F}}$ satisfies $(H\tilde{\mathcal{F}})$, with $\lambda(|s|, \Gamma) = \gamma^{\text{Lip}}$ and $\mu(|U|_\infty, \Gamma) \equiv 0$. Given the constant \bar{u} , we define $\mathcal{X}_\gamma = C[0, \bar{u}]$; our admissible set \mathcal{G} for this example is assumed to be a bounded subset of \mathcal{X}_γ , additionally satisfying

$$\mathcal{G} \subset \{\gamma \in \mathcal{X}_\gamma \mid \gamma(0) = 0, \gamma \text{ is nondecreasing, and } |\gamma(u_1) - \gamma(u_2)| \leq \gamma^{\text{Lip}}|u_1 - u_2|\}.$$

The assumptions we impose on γ are slightly weaker than those in [13].

3.A.1. Stefan problem. The Stefan problem (see, e.g., [10]) was originally formulated as a model for the melting of ice, although it is now used to model many phenomena (see, e.g., [10], [12], [29]). We consider the simplest one-dimensional one-phase model here. Let u represent the temperature of the water in the liquid phase and s represent the location of the ice-water interface. The units are such that the temperature in the ice phase is 0. The model equations can be written as

$$\begin{aligned}
 (3.4) \quad & \frac{\partial u}{\partial t} = \frac{\partial}{\partial x} \left(D \frac{\partial u}{\partial x} \right), \quad t \in (0, T], \quad x \in (0, s(t)); \\
 & \left. \frac{\partial u}{\partial x} \right|_{x=0} = g(t); \\
 & u(t, s(t)) = 0, \quad t \in (0, T]; \\
 & u(0, x) = u_0(x), \quad x \in [0, s_0]; \\
 & \frac{ds}{dt} = -\gamma D \frac{\partial u}{\partial x}(t, s(t)), \quad t \in [0, T]; \\
 & s(0) = s_0.
 \end{aligned}$$

Due to the pointwise evaluation of the flux, this $\tilde{\mathcal{F}}$ does not satisfy $(H\tilde{\mathcal{F}})$.

Our numerical method is essentially a Galerkin method, based on a weak formulation of the partial differential model equations. As can be seen in the proof of uniqueness (§ 4), for example, a natural measure for the solution $u(t, \cdot)$ of a parabolic equation

in weak form is given by the $H^1(0, 1)$ norm; thus the Stefan condition, involving $u_x(t, s(t))$, is too strong. After the change of variables to fixed domain, such a condition must be measured in the $W^{1,\infty}(0, 1)$ or $H^2(0, 1)$ topology (for almost every t). The majority of numerical methods for the Stefan problem, especially those accompanied by rigorous error analysis, are based on weaker formulations of the model equations—the enthalpy or variational formulations, for example. A survey of such methods can be found in [10].

3.A.2. Oxygen diffusion. We consider the second stage of the model presented in [21] for the diffusion of oxygen through tissue. As a precursor to radiation treatment for cancer, cells are saturated with oxygen. In the first stage of this procedure, oxygen is pumped into the cancerous tissue. In the second stage, the tissue is sealed and the treatment is applied. During the procedure, the oxygen diffuses, and the point of furthest penetration into the tissue recedes (hence the moving boundary). The procedure and the development of the simple model equations, which we present below, are discussed in [21]. Let u represent the concentration of oxygen in the tissue (after sealing) and s represent the furthestmost point of penetration. Our equations are

$$\begin{aligned}
 (3.5) \quad & \frac{\partial u}{\partial t} = D \frac{\partial^2 u}{\partial x^2} - \rho, \quad t \in (0, T], \quad x \in (0, s(t)); \\
 & \left. \frac{\partial u}{\partial x} \right|_{x=0} = 0; \\
 & u(t, s(t)) = 0, \quad t \in (0, T]; \\
 & u(0, x) = u_0(x), \quad x \in [0, s_0]; \\
 & \frac{\partial u}{\partial x}(t, s(t)) = 0, \quad t \in [0, T]; \\
 & s(0) = s_0.
 \end{aligned}$$

Here it is assumed that D and ρ are constants. Following Liapis, Lipscomb, and Crosser [21], we take the special form for the initial conditions

$$u_0(x) = \frac{\rho}{2D} (x - s_0)^2, \quad s_0 = \sqrt{\frac{2Du_i}{\rho}},$$

where, for convenience, we take $u_i = 1$. This particular example differs from all others presented here in that it is an “implicit” rather than an “explicit” moving boundary problem; i.e., rather than an equation for s or \dot{s} , we are given another condition on the state u at $x = s$. We have converted this (see [10, p. 166]) to the explicit condition

$$\frac{ds}{dt} = -\frac{D^2}{\rho} \frac{\partial^3 u}{\partial x^3}(t, s(t)).$$

Clearly, we require more smoothness on the state u than previously. Here, again, this form of $\tilde{\mathcal{F}}$ does not fit into our framework.

Remark (Notation). Note that for §§ 3.1–3.3, none of the parameters in $\tilde{\mathcal{F}}$ are spatially varying and so do not need to be redefined with the change of variables to fixed domain. We do not assume that this is the case in general; however, as it is the case for the specific examples we refer to throughout this paper, let us use the following notation: In all formulations of the problem, Γ will refer to the vector of parameters occurring in $\tilde{\mathcal{F}}$, with γ_i representing the components. For the biofilm example, $\Gamma = (\gamma_1, \gamma_2) \in C[0, \bar{u}] \times C[0, T]$; for the aerosomal process example, $\Gamma = (\gamma_1, \gamma_2) \in \mathbb{R}^2$; and for the polymer example, $\Gamma = \gamma \in C[0, \bar{u}]$.

4. Abstract formulation and approximations. It facilitates the development of the approximations and our convergence arguments to write our model equations in weak form (as in [2], [6]). At this point, it is necessary to distinguish between types of boundary conditions. Any essential boundary conditions must be homogeneous, in general, requiring a further change of variables. For the purpose of definiteness, yet trying to remain as general as possible in the exposition, we henceforth assume that $\alpha_{11} \equiv 0$, with α_{21} nonzero for all time. It should be clear how the following analyses would be modified for other types of boundary conditions. With $\alpha_{11} \equiv 0$, $\alpha_{12} \neq 0$, the left-hand boundary condition becomes $U(t, 0) = (-g(t))/\alpha_{12}$; let us replace $(-g(t))/\alpha_{12}$ by $g(t)$. Similarly, with $\alpha_{21} \neq 0$, we divide through the right-hand boundary condition, define $\alpha_2 \equiv \alpha_{22}/\alpha_{21}$, and replace $(h(t))/\alpha_{21}$ by $h(t)$. To obtain a homogeneous boundary condition, we define $\tilde{F}(t, y, g) = g(t)(y-1)^2$; the function \tilde{F} is constructed in such a way that the boundary condition in question becomes homogeneous while the other is unchanged (note that similar transformations could easily be defined for the cases where $\alpha_{21} \equiv 0$, $\alpha_{11} \neq 0$, or $\alpha_{11} = \alpha_{21} \equiv 0$). Let $H_B^1(0, 1) \equiv \{\phi \in H^1(0, 1) \mid \phi(0) = 0\}$, and define $V(t, y) = U(t, y) + \tilde{F}(t, y, g)$. Making this change of variables and writing the equation in weak form results in the following equation.

For each $t \in (0, T]$, we call $(V(t, \cdot), s(t)) \in H_B^1(0, 1) \times \mathbb{R}$ a solution if it satisfies, for all $\phi \in H_B^1(0, 1)$, the following equations:

$$\begin{aligned} \left\langle \frac{\partial V}{\partial t}, \phi \right\rangle &= -\frac{1}{s^2} \left\langle \mathcal{D} \frac{\partial V}{\partial y}, \frac{\partial \phi}{\partial y} \right\rangle - \frac{1}{s} \left\langle \mathcal{V}_1 V, \frac{\partial \phi}{\partial y} \right\rangle + \frac{\mathcal{V}_2}{s} \left\langle \frac{\partial V}{\partial y}, \phi \right\rangle + \frac{\dot{s}}{s} \left\langle y \frac{\partial V}{\partial y}, \phi \right\rangle + \langle \rho(V - \tilde{F}), \phi \rangle \\ &\quad + \langle \hat{F}, \phi \rangle + \frac{1}{s} (-\alpha_2 V(t, 1) + h(t)) \phi(1), \\ (4.1) \quad \frac{ds}{dt} &= \tilde{\mathcal{F}}(s, V - \tilde{F}(g); \Gamma), \\ V(0, y) &= V_0(y) \equiv U_0(y) + \tilde{F}(0, y, g), \\ s(0) &= s_0, \end{aligned}$$

where

$$\begin{aligned} (4.2) \quad \hat{F}(t, y, s, \dot{s}, \mathcal{D}, \mathcal{V}_1, \mathcal{V}_2, g, \dot{g}) &\equiv \tilde{F}_t + \frac{1}{s^2} \frac{\partial}{\partial y} (\mathcal{D} \tilde{F}_y) + \frac{1}{s} \frac{\partial}{\partial y} (\mathcal{V}_1 \tilde{F}) \\ &\quad + \frac{1}{s} \mathcal{V}_2 \tilde{F}_y + \frac{\dot{s}}{s} y \tilde{F}_y + F(t, y), \\ \tilde{F}(t, y, g) &= g(t)(y-1)^2. \end{aligned}$$

We now slightly redefine our parameter vector as $q \equiv (\mathcal{D}, \mathcal{V}_1, \mathcal{V}_2, \rho, \alpha_2, g, h, \Gamma, V_0, s_0)$. We define the parameter set $\mathbb{I} \equiv \{\alpha_2 \in C[0, T] \mid 0 \leq \alpha_2(t) \leq \bar{\alpha}_2\}$ and, as we require additional smoothness on the parameter g , we now assume that \mathcal{G} is a bounded subset of $C^1[0, T]$. We now write $\mathcal{Q} \equiv \mathcal{D} \times \mathfrak{B}_1 \times \mathfrak{B}_2 \times \mathfrak{P} \times \mathbb{I} \times \mathcal{G} \times \mathfrak{S} \times \mathcal{G} \times \mathcal{I}_v \times \mathcal{I}_s$.

We further assume any additional regularity of \mathcal{Q} necessary to ensure that the following hypotheses hold:

(HE) For any $q \in \mathcal{Q}$, a solution of (4.1) exists satisfying:

(HE1) $V \in C([0, T]; H_B^1(0, 1))$, and

(HE2) $s \in C^1(0, T)$, and $s(t) > 0$ for $t \in [0, T]$.

We assume the existence of solutions because the existence theory for moving boundary problems is incomplete. The biofilm model (§ 3.1) was derived and analyzed numerically in [8], but questions of existence of solutions to the model equations were not considered. In [26], the model for acrosomal elongation (§ 3.2) was formulated, and a perturbation analysis was used to examine qualitative properties of the solution; however, no existence results were proved. For a special case of the model equations of § 3.3 (D and σ constant, and smoother γ than we assume here), Fasano and Ricci [13] proved an existence result by showing the equivalence of the polymer model to a Stefan problem. In this case, they obtained stronger regularity of the solution than we assume above. As we show below, our parameter estimation techniques can be applied to a larger class of problems.

To ensure that the parameter estimation problem (\mathbb{P}) below is well posed, we prove that solutions of the weak equations, when they exist, are unique.

PROPOSITION 4.1. *Weak solutions of (4.1) are unique.*

The proof can be found in the Appendix.

For the development of our theoretical results in § 5, we require a compact parameter set $\mathcal{Q}_c \subset \mathcal{Q}$. A natural (in terms of our approximations, defined below) choice for a compact subset is as follows. Let $\Omega = [0, T] \times [0, 1]$, and let $\mathcal{Q}_c \equiv \mathcal{D}_c \times \mathcal{B}_{1c} \times \mathcal{B}_{2c} \times \mathcal{P}_c \times \mathcal{U}_c \times \mathcal{G}_c \times \mathcal{H}_c \times \mathcal{V}_c \times \mathcal{J}_{vc} \times \mathcal{J}_{sc}$, where

$$\mathcal{D}_c \equiv \{\mathcal{D} \in \mathcal{D} \cap W^{1,\infty}(\Omega) \mid |\mathcal{D}|_\infty \leq K_{d1}, |\mathcal{D}_t|_\infty \leq K_{d2}, |\mathcal{D}_y|_\infty \leq K_{d3}\},$$

$$\mathcal{B}_{1c} \equiv \{\mathcal{V} \in \mathcal{B}_1 \cap W^{1,\infty}(\Omega) \mid |\mathcal{V}|_\infty \leq K_{v1}, |\mathcal{V}_t|_\infty \leq K_{v2}, |\mathcal{V}_y|_\infty \leq K_{v3}\},$$

$$\mathcal{B}_{2c} \equiv \{\mathcal{V} \in \mathcal{B}_2 \mid |\mathcal{V}|_\infty \leq K_v\},$$

$$\mathcal{P}_c \equiv \{\rho \in \mathcal{P} \mid |\rho|_\infty \leq K_p\},$$

$$\mathcal{U}_c \equiv \{\alpha \in \mathcal{U} \cap W^{1,\infty}(0, T) \mid |\alpha|_\infty \leq K_{a1}, |\alpha'|_\infty \leq K_{a2}\},$$

$$\mathcal{H}_c \equiv \{h \in \mathcal{H} \cap W^{1,\infty}(0, T) \mid |h|_\infty \leq K_{h1}, |h'|_\infty \leq K_{h2}\},$$

$$\mathcal{J}_{vc} \equiv \{V_0 \in \mathcal{J}_v \cap W^{1,\infty}(0, 1) \mid |V_0|_\infty \leq K_{i1}, |V'_0|_\infty \leq K_{i2}\},$$

$$\mathcal{J}_{sc} \equiv \{s_0 \in \mathcal{J}_s \mid 0 < \underline{s} \leq s_0 \leq \bar{s}\}.$$

We assume that \mathcal{G}_c is a compact subset of $C^1[0, T]$. We must discuss \mathcal{G}_c in the context of particular examples; each of the remaining sets, as defined above, is compact in the $C(\delta)$ topology (where δ represents the relevant domain), and therefore the set \mathcal{Q}_c as defined is compact in the \mathcal{X}_q topology.

The parameter estimation problem can now be stated precisely as follows:

$$\begin{aligned} \text{Given } \tilde{J}(q) = & \sum_{i=1}^m \left\{ (\hat{s}_i - s(t_i; q))^2 \right. \\ (\mathbb{P}) \quad & \left. + (\hat{u}_i - s(t_i; q) \int_0^1 (V(t_i, y; q) - \tilde{F}(t_i, y; q)) dy)^2 \right\}, \\ \text{find Min } & \tilde{J}(q) \text{ subject to } (V(q), s(q)) \text{ the solution of (4.1).} \\ & q \in \mathcal{Q}_c \end{aligned}$$

The above problem must be discretized for the purposes of computation. We now describe our approximations; we follow, in spirit, ideas developed over several years. For just a sampling of references, see [2]–[6]. We begin by defining a finite-dimensional subspace of H_b^1 ; call it H^N . (In §§ 6 and 7, we are explicit about our choices of

approximation scheme; here, we keep our results as general as possible.) Let $P^N: H^0 \rightarrow H^N$ represent the orthogonal projection (in the H^0 topology) onto the finite-dimensional subspace. We require that the following conditions be satisfied:

(HA1) Given $\phi \in H_B^1$, $\|P^N\phi - \phi\| \rightarrow 0$ as $N \rightarrow \infty$,

(HA2) Given $\phi \in H_B^1$, $\left\| \frac{\partial}{\partial y} (P^N\phi - \phi) \right\| \rightarrow 0$ as $N \rightarrow \infty$.

For H^N the set of piecewise linear functions, and no Dirichlet boundary conditions (so that we replace H_B^1 by H^1), we use standard results from the theory of linear splines (see, e.g., [28]), combined with the dense inclusion of H^2 in H^1 and the fact that for $\phi \in H^1$, $\|P^N\phi - \phi\| \leq \|I^N\phi - \phi\|$ (where $I^N\phi$ represents the interpolant of ϕ in H^N), to argue that (HA1) and (HA2) hold (similar arguments appear in [5] and [19]). For cubic splines, the results are proved (for $\phi \in H^1$) in [5; Thm. A.5.4]. When we do have Dirichlet boundary conditions, we need to slightly modify the arguments described above; see, for example, [3]. Essentially, we rely on the fact that, for $\phi \in H_B^1$, $I_B^N\phi = I^N\phi$, where $I_B^N\phi$ represents the interpolant of ϕ in the spline subspace modified to obey the boundary condition.

Our approximate weak solution is the pair $(V^N(t, \cdot), s^N(t)) \in H^N \times \mathbb{R}$, which satisfies for each $t \in [0, T]$ and every $\phi \in H^N$

$$\begin{aligned} \left\langle \frac{\partial V^N}{\partial t}, \phi \right\rangle &= -\frac{1}{(s^N)^2} \left\langle \mathcal{D} \frac{\partial V^N}{\partial y}, \frac{\partial \phi}{\partial y} \right\rangle - \frac{1}{s^N} \left\langle \mathcal{V}_1 V^N, \frac{\partial \phi}{\partial y} \right\rangle + \frac{\mathcal{V}_2}{s^N} \left\langle \frac{\partial V^N}{\partial y}, \phi \right\rangle \\ &\quad + \frac{s^N}{s^N} \left\langle y \frac{\partial V^N}{\partial y}, \phi \right\rangle + \langle \rho(V^N - \tilde{F}), \phi \rangle + \langle \hat{F}, \phi \rangle \\ &\quad + \frac{1}{s^N} (-\alpha_2 V^N(t, 1) + h(t)) \phi(1), \end{aligned} \quad (4.1^N)$$

$$\frac{ds^N}{dt} = \tilde{\mathcal{F}}(s^N, V^N - \tilde{F}(g); \Gamma),$$

$$V^N(0, y) = P^N V_0(y),$$

$$s^N(0) = s_0.$$

As can be seen in § 6, the above system has a unique solution, which can be characterized as being the solution of a system of ordinary differential equations.

We must also discretize all infinite-dimensional parameter sets. Such “secondary” discretizations for parameter sets have been described in some detail elsewhere (see, for example, [4], [6], and [23]). We describe this only briefly here. For all of our variable parameters, we use linear spline approximations. Suppose that β is a function of one variable on the interval I (this is $[0, 1]$ or $[0, \bar{u}]$, etc.). Given M , we subdivide I uniformly and let S^M be the set of piecewise linear functions on that partition. A basis for S^M is the set of hat functions, denoted here by $\{b_i\}$. Any function in S^M then has the representation $\beta^M(x) = \sum_{i=1}^M \nu_i b_i(x)$. Now, instead of searching for β in some subset of $C(I)$, we search for the vector $\nu = (\nu_0, \nu_1, \dots, \nu_M) \in \mathbb{R}^{M+1}$.

Similarly, suppose that δ is a function of two variables on the rectangle $R = I_1 \times I_2$. Given $M = (M_1, M_2)$, we subdivide each of the intervals uniformly and let S^M be the set of piecewise bilinear functions on the partition. A basis for this space is the product of bases for each interval individually, i.e., $S^M(R) = S^{M_1}(I_1) \otimes S^{M_2}(I_2)$. Now $\delta(x, y)$

is approximated by $\delta^M = \sum_{i=1}^M \sum_{j=1}^{M_2} \mu_{ij} b_i(x) \tilde{b}_j(y)$, where $\{b_i\}$ is a basis for $S^{M_1}(I_1)$, and $\{\tilde{b}_j\}$ is a basis for $S^{M_2}(I_2)$. Thus δ^M is characterized by the vector $\mu = (\mu_{00}, \mu_{01}, \dots, \mu_{M_1 M_2}) \in \mathbb{R}^{(M_1+1)(M_2+1)}$.

Let $\mathcal{Q}_c^M \equiv I^M \mathcal{Q}_c$, where I^M designates the product interpolation operator (i.e., interpolate each component to the appropriate partition, leaving any constant parameters unchanged). Because each component interpolation operator is continuous, I^M is continuous (in the \mathcal{X}_q topology), and so \mathcal{Q}_c^M is compact. Moreover, for any $q \in \mathcal{Q}_c$, $I^M q \rightarrow q$ as $M \rightarrow \infty$.

Our corresponding approximate parameter estimation problem is stated as follows:

$$\begin{aligned} \text{Given } \tilde{J}^N(q) &= \sum_{i=1}^m \left\{ (\hat{s}_i - s^N(t_i; q))^2 + (\hat{u}_i - s^N(t_i; q) \int_0^1 (V^N(t_i, y; q) \right. \\ (\mathbb{P}^{N,M}) \quad &\quad \left. - \tilde{F}(t_i, y; q)) dy)^2 \right\}, \\ \text{find Min } \tilde{J}^N(q) &\text{ subject to } (V^N(q), s^N(q)) \text{ a solution of (4.1}^N\text{).} \\ &\quad q \in \mathcal{Q}_c^M \end{aligned}$$

Remark 4.1. As we compute with the transformed variable V when estimating the initial state, we will obtain $(\hat{V}_0^M, \hat{s}_0) \in H_B^1(0, 1) \times \mathbb{R}$; we can then construct the corresponding initial state estimate in the original variables as $\hat{u}_0(x) = \hat{V}_0^M(x/\hat{s}_0) - \hat{F}(0, (x/\hat{s}_0), \hat{g})$, where $\hat{u}_0 \in H^1(0, \hat{s}_0)$.

We make a final assumption, this one about the approximate boundary terms s^N . Note that for a fixed $\tilde{q} \in \mathcal{Q}$, we can bound $s(\tilde{q})$ and $\dot{s}(\tilde{q})$ (by virtue of hypothesis (HE2)); it is permissible that these bounds depend on \tilde{q} . We must assume that we can also bound the approximations in a similar way, but uniformly in N and q . Below, we state these hypotheses; it is convenient to combine the statements about the infinite-dimensional and approximate states into one:

Let $\tilde{q} \in \mathcal{Q}$ be fixed. Then

- (HN1) There exist positive constants $\underline{s}(\tilde{q})$, $\bar{s}(\tilde{q})$ such that for all $t \in [0, T]$, $\underline{s} \leq s(t; \tilde{q}) \leq \bar{s}$ and $\underline{s} \leq s^N(t; q) \leq \bar{s}$ for all N and for any $q \in \mathcal{Q}$;
- (HN2) There exists a positive constant $K_s(\tilde{q})$ such that for all $t \in [0, T]$, $|\dot{s}(t; \tilde{q})| \leq K_s$ and $|\dot{s}^N(t; q)| \leq K_s$ for all N and for any $q \in \mathcal{Q}$.

This constant K_s is now understood to be the same as that used in the definition of the set \mathcal{B}_2 . The constants \underline{s} and \bar{s} are now understood to be those in the definition of the set \mathcal{I}_{sc} .

While physical or biological principles will usually dictate that the conditions of (HN1) and (HN2) hold for the “original” (infinite-dimensional) model equations, these principles do not extend to the approximations. For some example problems, the hypotheses for the approximations can be assured by defining appropriate constraints within the parameter space \mathcal{G} . In some cases, however, it may be necessary to alter the definition of the approximation scheme. For the biofilm example, as expressed in (3.1) and the glassy polymers example (3.3), an appropriate definition of the set \mathcal{G} will ensure that (HN1) and (HN2) hold. For the acrosomal elongation example (3.2), we need to make a slight change in the approximations.

Consider the biofilm example. Since s^N satisfies an equation of the form $\dot{s}^N(t) = s^N(t)M(t)$, the above hypotheses will be satisfied as long as s_0 is sufficiently restricted and M is bounded. Given that \mathcal{G} is a bounded subset of \mathcal{X}_γ (as defined in § 3.1), and s_0 belongs to a bounded subset of \mathbb{R} ; the conclusion follows. Similarly, for the polymer example of § 3.3, the definition of \mathcal{G} and the fact that \mathcal{G} and \mathcal{I}_s are bounded subsets of \mathcal{X}_γ and \mathbb{R} ensure that s^N will satisfy (HN1) and (HN2).

Now consider the model for acrosomal elongation. If the model is realistic, then it is reasonable to assume that for $\tilde{q} \in \mathcal{Q}$, $0 \leq \dot{s}(\tilde{q}) \leq K_s$ for some constant $K_s = K_s(\tilde{q})$. This then implies that statement (HN1) holds for $s(\tilde{q})$, and also that $\tilde{\gamma}_2/\tilde{\gamma}_1 \leq U(t, 1; \tilde{q}) \leq (K_s + \tilde{\gamma}_2)/\tilde{\gamma}_1$. We can hold no such expectations for the approximation s^N , however. We therefore define a “constraint projection” operator p_Γ for any $\Gamma \in \mathcal{G}$ by

$$\begin{aligned} p_\Gamma(\theta) &= \frac{\gamma_2}{\gamma_1} \quad \text{if } \theta \leq \frac{\gamma_2}{\gamma_1}, \\ p_\Gamma(\theta) &= \theta \quad \text{if } \frac{\gamma_2}{\gamma_1} \leq \theta \leq \frac{\gamma_2 + K_s}{\gamma_1}, \\ p_\Gamma(\theta) &= \frac{\gamma_2 + K_s}{\gamma_1} \quad \text{if } \theta \geq \frac{\gamma_2 + K_s}{\gamma_1}. \end{aligned}$$

While we have no guarantees on $U^N(t, 1)$ and therefore no guarantees on s^N , we can be sure that for any Γ , $p_\Gamma(U^N(t, 1))$ satisfies $0 \leq \gamma_1 p_\Gamma(U^N(t, 1)) - \gamma_2 \leq K_s$; thus, we define our approximate dynamics for s as $ds^N/dt = \tilde{\mathcal{F}}(s^N, p_\Gamma(U^N); \Gamma)$. With s^N defined in this way, we are assured that (HN1) and (HN2) hold. Moreover, as guaranteed by the lemma below, we have not destroyed the convergence properties of our approximation scheme. Note that we are to perform our computations with the quantity V^N , not U^N ; however, in this particular example, $\tilde{\mathcal{F}}$ requires $V^N(t, 1)$, which (by construction of the transformation taking U to V) is equal to $U^N(t, 1)$. Since it is conceivable that a general $\tilde{\mathcal{F}}$ might involve more than just a boundary value, we assume henceforth that when p_Γ is necessary, our approximate dynamics for s are constructed as, for $g \in \mathcal{G}$ and $\Gamma \in \mathcal{G}$,

$$(4.3) \quad \frac{ds^N}{dt} = \tilde{\mathcal{F}}(s^N, p_\Gamma(V^N - \tilde{F}(g)); \Gamma).$$

For the example of the elongation of the acrosomal process described above, we can easily prove the following lemma.

LEMMA 4.1. *Given $\Gamma \in \mathcal{G}$, let p_Γ be the “projection” defined above. Then*

(P1) *For each $\Gamma \in \mathcal{G}$ and $\theta_1, \theta_2 \in \mathbb{R}$, $|p_\Gamma(\theta_1) - p_\Gamma(\theta_2)| \leq |\theta_1 - \theta_2|$.*

(P2) *For $\tilde{\Gamma} \in \mathcal{G}$, let $\tilde{V} \in L^\infty(0, 1)$ be such that $p_{\tilde{\Gamma}}(\tilde{V}) = \tilde{V}$ (i.e., \tilde{V} satisfies the constraints corresponding to $\tilde{\Gamma}$). For any $\{\Gamma^k\} \subset \mathcal{G}$, with $\Gamma^k \rightarrow \tilde{\Gamma}$, $|p_{\Gamma^k}(\tilde{V}) - p_{\tilde{\Gamma}}(\tilde{V})|_\infty \equiv |p_{\Gamma^k}(\tilde{V}) - \tilde{V}|_\infty \rightarrow 0$ as $\Gamma^k \rightarrow \tilde{\Gamma}$.*

In general, we assume either that (HN1) and (HN2) holds directly (i.e., via a suitable restriction on the parameter set \mathcal{G}) or, as in the example above, that p_Γ can be constructed so that (HN1) and (HN2) hold, while satisfying (P1) and (P2).

5. Convergence arguments. In this section we state our theoretical results, with the proofs of Theorems 5.3 and 5.4 appearing in the Appendix.

As can be seen in § 6, (V^N, s^N) , and hence \tilde{J}^N , is continuous in the parameters. Since each \mathcal{Q}_c^M is compact, each problem $(\mathbb{P}^{N,M})$ therefore has a solution. Thus, for increasing N and M , we can solve each of these finite-dimensional problems, generating a sequence of best-fit parameters $\{\hat{q}^{N,M}\}$.

Our key result is that for any arbitrary convergent sequence of parameters $q^M \rightarrow \tilde{q}$ in \mathcal{X}_q with $M \rightarrow \infty$, the solution of the finite-dimensional equations (4.1^N) , solved with the parameter q^M , converges as $M, N \rightarrow \infty$ to the solution of (4.1) using \tilde{q} . This result is key because it allows the proof of the following two theorems. The arguments used are slight modifications of those appearing elsewhere (see, e.g., [3]), and so we describe them only briefly.

THEOREM 5.1. *Suppose that $\mathcal{Q}_c^M, \mathcal{Q}_c$ are all compact in the \mathcal{X}_q topology, with $M \rightarrow \infty$. Assume that for any sequence $\{q^M\} \subset \mathcal{Q}$, with $q^M \rightarrow \tilde{q} \in \mathcal{Q}_c$, it follows that $(V^N(q^M), s^N(q^M)) \rightarrow (V(\tilde{q}), s(\tilde{q}))$ in $H^0(0, 1) \times \mathbb{R}$ as $N, M \rightarrow \infty$ for each $t \in [0, T]$.*

Then $\{\hat{q}^{N,M}\}$ has a subsequence $\{\hat{q}^{N_k, M_k}\}$, with $\hat{q}^{N_k, M_k} \rightarrow q^$ where $q^* \in \mathcal{Q}_c$ minimizes \tilde{J} .*

In fact, in the case where problem (\mathbb{P}) has a unique solution, we can argue full sequential convergence of the approximate parameter estimates. The proof of this theorem relies heavily on the compactness assumptions on the parameter sets, standard spline estimates, and the convergence statement that we have identified as being our key result. We outline the arguments (for details of similar arguments, see [4], [6], and [23]).

For the sets $\mathfrak{B}_{2c}, \mathfrak{B}_c, \mathfrak{U}_c, \mathfrak{S}_c, \mathcal{I}_{vc}$, and any of the sets \mathcal{G}_c considered in §§ 3.1–3.3, we can characterize P_c^M (where P represents any of the above sets) as $P_c^M = S^M \cap P_c$ (S^M represents the appropriately defined linear spline subspace). Then, given $\{\hat{p}^{N,M}\} \subset P_c^M \subset P_c$, we can directly extract a convergent subsequence $\{\hat{p}^{N_k, M_k}\}$ with $\hat{p}^{N_k, M_k} \rightarrow p^*$ in P_c . For the remaining parameter sets, this characterization does not hold, but we can for any $\hat{p}^{N,M} \in P_c^M$ (where P now represents one of the second group of sets) identify $\bar{p}^{N,M} \in P_c$ such that $\hat{p}^{N,M} = \iota^M \bar{p}^{N,M}$ (we use ι^M to designate the interpolation operator from P_c to P_c^M). From the sequence $\{\bar{p}^{N,M}\} \subset P_c$ then, we extract a convergent subsequence, with $\bar{p}^{N_k, M_k} \rightarrow p^*$ in P_c . Standard spline estimates (in particular, the fact that for $p \in P_c$, $\iota^M p \rightarrow p$ as $M \rightarrow \infty$) can then be used to also show that $\hat{p}^{N_k, M_k} \rightarrow p^*$. (In the case of estimating g within a Dirichlet boundary condition, the need for C^1 compactness requires that extra conditions be imposed on the set \mathfrak{G}_c^M).

Given the subsequential convergence of the parameters, it is then easy to apply the key result, together with the fact that $I^M q \rightarrow q$ as $M \rightarrow \infty$ for any $q \in \mathcal{Q}_c$, and obtain the stated conclusion.

While it may be unreasonable to expect problems $(\mathbb{P}^{N,M})$ and (\mathbb{P}) to have unique solutions, our compactness assumptions can be used to ensure that our method is stable in the following sense (see [2], [5]). Let $\tilde{q}^{N,M}(\hat{u}^k, \hat{s}^k)$ represent the set of all solutions to problem $(\mathbb{P}^{N,M})$ obtainable with the set of data (\hat{u}^k, \hat{s}^k) ; let $\tilde{q}(\hat{u}^0, \hat{s}^0)$ represent the set of all solutions to problem (\mathbb{P}) obtainable with the data (\hat{u}^0, \hat{s}^0) . Then, following [2], we call our approximation method stable if $\text{dist}(\tilde{q}^{N,M}(\hat{u}^k, \hat{s}^k), \tilde{q}(\hat{u}^0, \hat{s}^0)) \rightarrow 0$ as $N, M \rightarrow \infty$, and $(\hat{u}^k, \hat{s}^k) \rightarrow (\hat{u}^0, \hat{s}^0)$ in \mathbb{R}^{2m} , where “dist” represents the Hausdorff distance between two sets.

THEOREM 5.2. *Assume that the hypotheses of Theorem 5.1 hold. Then our approximation scheme is stable in the sense defined above.*

The proof of this theorem involves arguments similar to those of Theorem 5.1, with the additional observation that \tilde{J}^N depends continuously on the data.

We now turn to our main theorem. We are interested in the quantities $(V(\tilde{q}), s(\tilde{q}))$ and $(V^N(q^M), s^N(q^M))$, which represent the solution of (4.1) with the parameter \tilde{q} , and the solution of (4.1^N) with the parameter q^M , respectively. To reduce notation, we henceforth write (V, s) for $(V(\tilde{q}), s(\tilde{q}))$, and $(V^{N,M}, s^{N,M})$ for $(V^N(q^M), s^N(q^M))$.

THEOREM 5.3. *Assume that the hypotheses (HE1), (HE2), (H \mathcal{F}), (HA1), (HA2), and (HN1), (HN2) hold (and, if necessary, (P1) and (P2)), and that $\{q^M\}$ is an arbitrary sequence of parameters in \mathcal{Q} with $q^M \rightarrow \tilde{q} \in \mathcal{Q}$. Then it follows that for each $t \in [0, T]$, $\|V - V^{N,M}\|^2 + |s - s^{N,M}|^2 \rightarrow 0$, as $N, M \rightarrow \infty$.*

We must consider the equations modelling the diffusion in polymers (see (3.3)) as a separate case; here we have $\alpha_2 = \alpha_2(\dot{s}) = \dot{s}$ and $h = h(\dot{s}) = -\sigma\dot{s}$. We have no parameter set \mathfrak{U} , and our set \mathfrak{S}_c will be defined as $\mathfrak{S}_c \equiv \{\sigma \in W^{1,\infty}[0, T] | 0 < \sigma \leq \sigma(t) \leq \bar{\sigma} \text{ and } |\sigma'|_\infty \leq K_\sigma\}$. We rely on the specific form of the functional $\tilde{\mathcal{F}}$; for the polymer

example, $\tilde{\mathcal{F}} = \gamma(U(t, 1)) \equiv \gamma(V(t, 1))$, where we assume $\gamma \in \mathcal{G}_c \equiv \{\gamma \in \mathcal{G} \mid \|\gamma\|_\infty \leq K_\gamma\}$, and \mathcal{G} was defined in § 3.3. We rely on the fact that $V(t, 1) = U(t, 1) \geq 0$, since (see [13]) u and therefore U are nonnegative for all (t, x) , (t, y) , respectively.

THEOREM 5.4. *Assume that the hypotheses of Theorem 5.3 hold with, additionally, $V(t, 1) \geq 0$ for all $t \in [0, T]$, and \mathcal{G}_c , \mathcal{E}_c , defined above. Then, for any $q^M \rightarrow \tilde{q}$ in \mathcal{Q} , it follows that the solutions of the polymer equations (3.3) converge in the sense of the conclusion of Theorem 5.3.*

6. Numerical implementation. We describe the implementation of our estimation procedure, beginning with the conversion of the abstract approximate equations (4.1^N) to an equivalent system of ordinary differential equations. We typically use either linear or cubic splines. If our underlying state space is $H^1(0, 1)$, a convenient basis for H^N when using linear splines is the set of “hat” functions, and for cubic splines is the set of cubic B -splines (as defined, e.g., in [27]). It is straightforward to modify the set of basis elements to satisfy essential boundary conditions (when the state space is H_B^1).

For fixed N , consider the approximate weak equations (4.1^N). Let \mathbb{N} be the dimension of H^N , and let $\{B_i\}_{i=1}^{\mathbb{N}}$ be a basis for this subspace. Since $V^N(t, \cdot) \in H^N$ for each t , we can write $V^N(t, \cdot) = \sum_{i=1}^{\mathbb{N}} w_i(t) B_i$. Let $w(t) = (w_1(t), w_2(t), \dots, w_{\mathbb{N}}(t))$, and let us write s for s^N . Calculating $(\partial V^N)/\partial t$ and $(\partial V^N)/\partial y$, and substituting $\phi = B_i$ for $i = 1, 2, \dots, \mathbb{N}$ into (4.1^N), we obtain the following system of ordinary differential equations in w and s :

$$\begin{aligned} M\dot{w} = & -\frac{1}{s^2} K(\mathcal{D})w - \frac{1}{s} L^1(\mathcal{V}_1)w + \frac{\mathcal{V}_2(s)}{s} L^2w + \frac{\dot{s}}{s} Tw \\ & + p(\rho, w) + f(s, \dot{s}; q) + r(s, w; \alpha_2, h), \\ \dot{s} = & \tilde{\mathcal{F}}\left(s, \sum_{i=1}^{\mathbb{N}} w_i B_i - \tilde{F}(g); \Gamma\right), \end{aligned} \quad (6.1)$$

with initial conditions $Mw(0) = c$, $s(0) = s_0$; we have defined the following matrices:

$$\begin{aligned} M_{ij} &= \langle B_i, B_j \rangle, \quad [K(\mathcal{D})]_{ij} = \langle \mathcal{D}B'_i, B'_j \rangle, \quad [L^1(\mathcal{V}_1)]_{ij} = \langle \mathcal{V}_1 B'_i, B'_j \rangle, \\ L_{ij}^2 &= \langle B_i, B'_j \rangle, \quad \text{and} \quad T_{ij} = \langle y B_i, B'_j \rangle, \end{aligned}$$

and we have defined the following vectors:

$$\begin{aligned} p_i &= \langle \rho(w), B_i \rangle, \quad f_i = \langle \hat{F}(s, \dot{s}; q), B_i \rangle, \quad r_i = \frac{1}{s} \left(-\alpha_2 \sum_{j=1}^{\mathbb{N}} w_j B_j(1) + h \right) B_i(1), \\ &\text{and} \quad c_i = \langle V_0, B_i \rangle. \end{aligned}$$

As described in § 4, we also discretize the parameter spaces. In the examples presented below, we have identified diffusion coefficients of the form $\mathcal{D} = \mathcal{D}(t)$ or $\mathcal{D} = \mathcal{D}(y)$. In the first case, we would approximate \mathcal{D} by $\mathcal{D}^M(t) = \sum_{k=1}^M d_k b_k(t)$ (see § 4), and then at each t we need to evaluate $[K(\mathcal{D})]_{ij} = (\sum_{k=1}^M d_k b_k(t)) \langle B'_i, B'_j \rangle$; the unknown parameter is the vector $d = (d_1, d_2, \dots, d_M) \in \mathbb{R}^M$. In the second case, we would approximate \mathcal{D} by $\mathcal{D}^M(y) = \sum_{k=1}^M d_k b_k(y)$, in which case we need to evaluate $[K(\mathcal{D}^M)]_{ij} = \sum_{k=1}^M d_k \langle b_k B'_i, B'_j \rangle$; again, the unknown is the vector $d \in \mathbb{R}^M$. While we have not estimated coefficients of the form $\mathcal{D} = \mathcal{D}(t, y)$ here, the implementation would be similar to that in diffusion equations on a fixed domain, as treated in [4]. Similar comments apply to the estimation of variable \mathcal{V}_1 .

In the approximation for ρ , we must first estimate the constant \bar{u} , which determines the interval on which to define the approximation. This is a problem that arises whenever

we approximate an unknown nonlinearity, and it has been discussed in [6] and [23]. As reported there, it is our experience that we can make a reasonable guess for an appropriate \bar{u} , then correct during the estimation procedure. Similar comments apply to the determination of K_s for the estimation of the state-dependent coefficient \mathcal{V}_2 . Unfortunately, (more so when using a nonvector computer), for each t the vector p must be recomputed.

The vector f can be written as

$$f_i = \dot{g} \langle \psi_1, B_i \rangle - \frac{g}{s^2} \langle \mathcal{D} \psi_2, B_i \rangle - \frac{g}{s} \langle \mathcal{V}_1 \psi_1, B_i \rangle + \frac{g \mathcal{V}_2(s)}{s} \langle \psi_2, B_i \rangle + \frac{g \dot{s}}{s} \langle \psi_3, B_i \rangle + \langle F, B_i \rangle,$$

where the ψ_i are known functions of y (see (4.2)). The parameters \mathcal{D} and \mathcal{V}_1 would be handled as described above. We would approximate g by $g^{\mathcal{M}'} = \sum_{k=1}^{\mathcal{M}'} g_k \beta_k(t)$, giving $\dot{g}^{\mathcal{M}'} = \sum_{k=1}^{\mathcal{M}'} g_k \dot{\beta}_k(t)$, thus reducing the estimation of g to the estimation of a vector in $\mathbb{R}^{\mathcal{M}'}$. In the vector r , the parameters α_2 and h may both be time dependent, and would be approximated in the same manner as g .

The unknowns appearing in the functional $\tilde{\mathcal{F}}$ are, in all cases we consider here, either constants or functions of the state u . In the latter case, the approximations are then analogous to that for ρ .

Since for a given approximation scheme, (4.1^N) and (6.1) are equivalent, we can address the question raised in § 4 of existence and uniqueness of solutions, and justify our statement in § 5 that J^N is continuous in the parameters by appealing to the theory of ordinary differential equations.

Our parameter set is bounded and, for any parameters, s^N is a priori bounded in $[\underline{s}, \bar{s}]$ (this is ensured by construction; see (HN1), (HN2), and associated comments). Moreover, the dynamics for w are linear in w except for the term involving ρ ; by assumption on the parameter sets, ρ is Lipschitz and satisfies $|\rho|_\infty \leq K_\rho$ for some constant K_ρ . Thus we can argue the existence of a unique solution by observing that the dynamics of (6.1) satisfy a local Lipschitz condition. It is easy to argue that the solution exists for all $t \in [0, T]$.

The continuity of solutions on parameters also follows from a straightforward argument; rewriting (6.1) as a system in $W = (w, s)$, $\dot{W} = F(W)$, it is easy to see that F depends continuously on the parameters, from which it follows that the solutions do, as well.

To obtain a solution to $(\mathbb{P}^{N,M})$, we proceed as follows. For a fixed N and M , we compute and store as many of the entries of the various matrices and vectors as possible. All of the numerical quadratures required for all the matrices need be computed only once for a given N . All of the numerical quadratures required to form c , and all for f , except possibly those involving the term F (note that in (3.1)–(3.2), $F \equiv 0$), need be computed only once for a given N . Forming the vector p is the only computationally intensive operation (we are currently using a scalar machine; on a vector computer, this would not be the case).

Given an initial guess q^0 for the unknown parameters, we solve system (6.1) for $(w(q^0), s(q^0))$, obtaining values at the set of times $\{t_i\}_{i=1}^m$. This solution is hastened by the banded structure of the matrices (all matrices are three-banded for linear splines, seven-banded for cubics). We then evaluate $\tilde{J}^N(q^0)$, which involves (see $(\mathbb{P}^{N,M})$) the term

$$\int_0^1 (V^N(t_i, y; q^0) - \tilde{F}(t_i, y; q^0)) dy = \sum_{k=1}^N w_k(t_i; q^0) \int_0^1 B_k dy - g^0(t_i) \int_0^1 (y-1)^2 dy.$$

Note that the evaluation of \tilde{J}^N is a simple computation involving the data s , w , and

g , and a set of constants that depend on the basis, and therefore are computed once for a given N . This process is then repeated iteratively until convergence to $\hat{q}^{N,M}$ is obtained. We use the IMSL subroutine ZXSSQ (based on the Levenberg-Marquardt algorithm) to perform the minimization over \mathcal{Q}_c^M . For a given q^i , the solution of system (6.1) is performed using the IMSL routine DGEAR (based on Gear's method).

7. Numerical test examples. Here we present the results of some test examples, which were motivated by the applications described in § 3. We also give examples of the models (§§ 3.A.1 and 3.A.2) that do not satisfy (H \mathcal{F}). For each example, we have tried to follow these governing equations as closely as possible; however, we have had to make various minor changes (which we indicate below) to obtain analytical solutions.

In each of Examples 1-3, below, we choose "true" parameters and a "true" solution, which satisfies the boundary conditions and the equation $\dot{s} = \tilde{\mathcal{F}}$; we then evaluate $F(t, y)$ so that the partial differential equation for U is satisfied (see (2.2)). Our "data" is then obtained by choosing the set of times $\{t_i\}_{i=1}^m$, and then evaluating $\hat{s}_i = s(t_i)$, and $\hat{u}_i = s(t_i) \int_0^1 U(t_i, y) dy$ (or, alternatively, $\hat{s}_i = s(t_i) - s_0$ or, e.g., $\hat{u}_i = U(t_i, 1)$).

We have unless otherwise specified used cubic splines to approximate the state U^N (or V^N), with $N = 6$, and linear splines to approximate variable parameters (with various values of M , as given below). All of the computations have been performed on the IBM 4381 at the University of North Carolina.

Example 1. In our first set of examples, we estimate parameters within the simplified model equations describing diffusion of a contaminant through a biofilm layer (3.1), transformed to the fixed domain, as follows:

$$\begin{aligned} \frac{\partial U}{\partial t} &= \frac{1}{s^2} \frac{\partial}{\partial y} \left(\mathcal{D} \frac{\partial U}{\partial y} \right) + \frac{\dot{s}}{s} y \frac{\partial U}{\partial y} + \rho(U) + F(t, y), & y \in (0, 1), \quad t \in (0, T]; \\ \left(\frac{\mathcal{D}}{s} \frac{\partial U}{\partial y} \right) \Big|_{y=0} &= g(t); \\ \left(\frac{\mathcal{D}}{s} \frac{\partial U}{\partial y} \right) \Big|_{y=1} + \alpha_2 U(t, 1) &= h(t), \quad t \in (0, T]; \\ U(0, y) &= U_0(y), \quad y \in [0, 1]; \\ \frac{ds}{dt} &= s(t) \int_0^1 (\gamma_1(U(t, y)) - \gamma_2(t)) dy, \quad t \in [0, T]; \\ s(0) &= s_0. \end{aligned}$$

We present the results of three estimation problems with variations of the above equations. As described above, to generate our data we choose the true parameters and a true solution (U, s) , which satisfies the boundary conditions and the equation $\dot{s} = \tilde{\mathcal{F}}(s, U; \gamma)$; we then compute

$$F(t, y) = \frac{\partial U}{\partial t} - \frac{1}{s^2} \frac{\partial}{\partial y} \left(\mathcal{D} \frac{\partial U}{\partial y} \right) + \frac{\dot{s}}{s} y \frac{\partial U}{\partial y} + \rho(U).$$

This term F then is not part of the model but is used to obtain an analytical solution. Also, while in the model of [8] there is initially no substrate in the biofilm layer (so $U_0 = 0$), we have a nonzero initial condition. This change is also made to facilitate the construction of an analytical solution.

In [24], linearized dynamics for the diffusion process were considered (i.e., $\rho \equiv 0$; this makes the necessary computations much faster). Within this simpler model, an unknown diffusion coefficient (both time dependent and spatially varying) was estimated. There, data of the form $\hat{u}_i = u(t_i, 0)$ was used. This would correspond, however, to observations of substrate concentration at the interface between the carbon particle and the biofilm layer; while this may be of mathematical interest, it is probably not likely in a true experiment. It is more likely that we could obtain measurements of substrate concentration at the outer edge of the biofilm layer (i.e., the data would be $\hat{u}_i \sim u(t_i, s(t_i))$) or measurements of total concentration within the biofilm layer (then the data would be $\hat{u}_i \sim \int_0^{s(t_i)} u(t_i, x) dx$). Our first two examples, below, repeat some of those estimations, using these more realistic data.

Example 1A. We take $\rho \equiv 0$, and estimate \mathcal{D} , with the true parameter given by $\mathcal{D}(y) = e^{-y}$. We fix all other parameters as follows: $g(t) = e^{-2t}$, $\alpha_2 = 1$, $h(t) = e^{-2t} - 1 + e^{-t+1}$, $\gamma_1(\theta) = \theta/(1+\theta)$, $\gamma_2(t) = e^t(e^{-1} - 1)$. Our true solution is chosen to be $U(t, y) = e^{-t+y} - 1$, $s(t) = e^t$; we then compute $F(t, y) = -e^{-t+y}(1+y)$ (this is assumed to be known). Our data is chosen of the form $\hat{u}_i \sim u(t_i, s(t_i)) = U(t_i, 1)$, with ten values of time chosen from the interval $[0, 1.5]$. We use $M = 5$ for the approximation of \mathcal{D} , obtaining the best fit parameter in Fig. 1(a). As can be seen in the figure, our initial guess is $\mathcal{D}^0 \equiv 1$. We obtain a residual of $\tilde{J}^6(\mathcal{D}^{6,5}) = 0.344 \times 10^{-10}$.

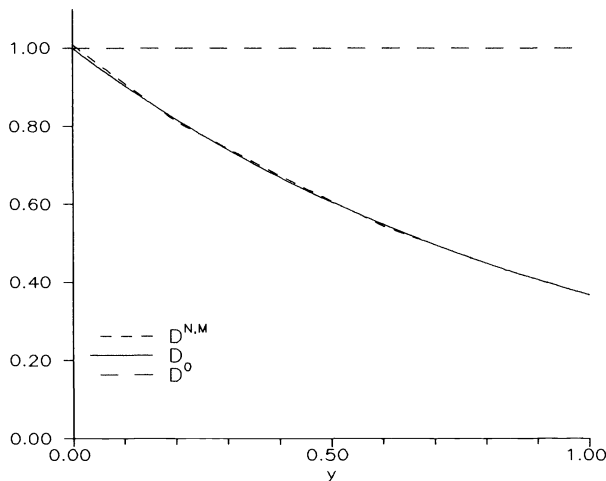
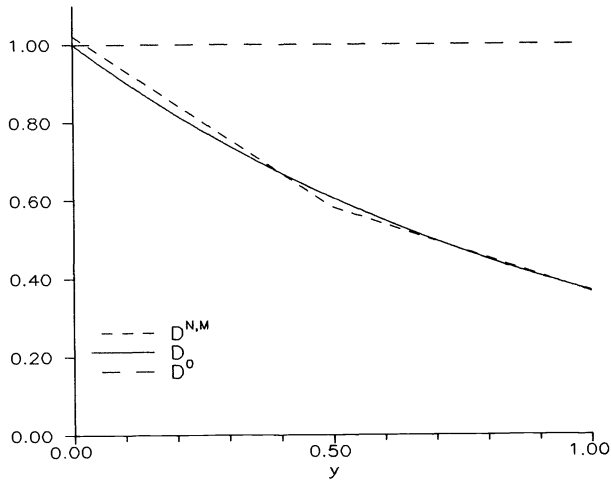
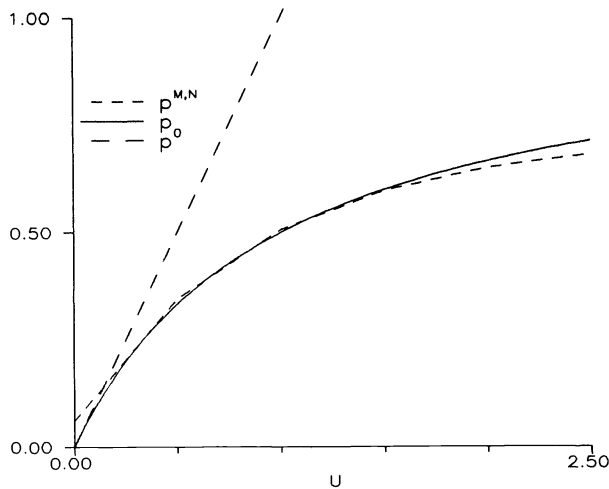


FIG. 1(a). Estimation of $D(y)$, $M = 5$.

Example 1B. This example is the same as Example 1A, except that we use the other type of realistic data; our data here is $\{\hat{u}_i\}_{i=1}^{10}$, where $\hat{u}_i \sim \int_0^{s(t_i)} u(t_i, x) dx = s(t_i) \int_0^1 U(t_i, y) dy$, with ten times chosen from the interval $[0, 1.5]$. Our results from the estimation are presented in Fig. 1(b). Our initial guess for the diffusion parameter is $\mathcal{D}^0 \equiv 1$. Using $M = 2$, we obtain the best fit parameter $\mathcal{D}^{6,2}$ graphed in the figure. Our residual is $\tilde{J}^6(\mathcal{D}^{6,2}) = 0.176 \times 10^{-9}$.

Example 1C. For this example, we use the fully nonlinear problem, and estimate both ρ and γ_1 . These parameters are assumed to be equal, so that only one parameter is estimated, but the unknown parameter appears in both places of the model equations (the biology is such that we expect these parameters to be proportional). The true parameter is taken to be $\rho(\theta) = \gamma_1(\theta) = \theta/(1+\theta)$ (we note that in [8] this function is parametrized as $k\theta/(K+\theta)$). We fix all remaining parameters as follows: $\mathcal{D} \equiv 1$,

FIG. 1(b). Estimation of $D(y)$, $M=2$.FIG. 1(c). Estimation of $p(u)$, $M=5$.

$g(t) = e^{-2t}$, $\alpha_2 = 1$, $h(t) = e^{1-2t} + e^{1-t}$, and $\gamma_2(t) = -1 + \ln((1 + e^{1-t})/(1 + e^{-t}))$. Our true solution is $U(t, y) = e^{-t+y}$, $s(t) = e^t$, with F (assumed known) then given by $F(t, y) = -(e^{-t} + e^{-3t})e^y - y e^{-t+y} + e^{-t+y}/(1 + e^{-t+y})$. Our data is of the form $\hat{u}_i \sim U(t_i, 1)$, with ten time values chosen from the interval $[0, 2.5]$. We choose $\bar{u} = 2.5$ (see discussion in § 6), and estimate $\rho = \gamma_1$ over the interval $[0, 2.5]$ with $M=5$. The results of this example are presented in Fig. 1(c). We obtain $\tilde{J}^6(\rho^{6,5}, \gamma_1^{6,5}) = 0.163 \times 10^{-6}$, having begun with the initial guess $\rho^0(\theta) \equiv \gamma_1^0(\theta) = \theta$.

In Example 1A, approximately 450 evaluations of the function J^N are required, with a total CP time of approximately 90 minutes. Thus the average time-per-function evaluation is about 0.2 min/FE. Approximately 850 function evaluations are required, with a total of approximately 300 CP minutes for Example 1C, where we have now included the nonlinear term. This is an average of about 0.31 min/FE.

Example 2. This example is chosen to represent the elongation of the acrosomal process, as described by (3.2). The model equations (transformed to fixed domain and

then to homogeneous Dirichlet boundary condition) are

$$\frac{\partial V}{\partial t} = \frac{\mathcal{D}}{s^2} \frac{\partial^2 V}{\partial y^2} + \frac{\dot{s}}{s} (y-1) \frac{\partial V}{\partial y} + \hat{F}(t, y, s, \dot{s}, \mathcal{D}, g, \dot{g}), \quad t \in (0, T], \quad y \in (0, 1);$$

$$V(t, 0) = 0;$$

$$\frac{\mathcal{D}}{s} \frac{\partial V}{\partial y}(t, 1) + \alpha(t) V(t, 1) = h(t), \quad t \in (0, T];$$

$$V(0, y) = V_0(y), \quad y \in [0, 1];$$

$$\frac{ds}{dt} = \gamma_1 V(t, 1) - \gamma_2, \quad t \in [0, T];$$

$$s(0) = s_0.$$

We set $k_1 = k_{-1} = 1$. We replace u_0 (a constant in the model of [26]) by the time-dependent function $g(t)$ (it is then no longer equal to the initial condition for u , which is also nonconstant here), and we add a nonhomogeneous term F , all to obtain an analytical solution.

In this example, we estimate the boundary terms α and h (we assume that they are equal; as can be seen in the model derivation of [26], they should be proportional). The true parameter is chosen to be $\alpha(t) = h(t) = 5e^{-t}$. The other parameters have all been fixed at $\mathcal{D} \equiv 2$, $g(t) = 1/(1+t)^2$, $\gamma_1 = 1$, and $\gamma_2 = 1$. The true solution is chosen to be $s(t) = 2 - t^2/(1+t)$ and

$$U(t, y) = \frac{1}{(1+t)^2} + \frac{5e^{-t}}{2} y(y-1) \left(2 - \frac{t^2}{1+t} \right) \left(\frac{1}{(1+t)^2} - 1 \right),$$

and then F (assumed known) is calculated as $F(t, y) = U_t - (\mathcal{D}/s^2)U_{yy} - (\dot{s}/s)(y-1)U_y$ (an explicit expression is very unwieldy and not of much interest). Finally, we calculate

$$\hat{F} = \frac{4}{(s^N)^2} \frac{1}{(1+t)^2} + 2 \left(\frac{\dot{s}^N}{s^N} + \frac{1}{(1+t)} \right) \frac{(y-1)^2}{(1+t)^2} + F(t, y).$$

Note that in the calculation of F , we use the "true" (known functions) s , U , while in \hat{F} we compute with the current approximation.

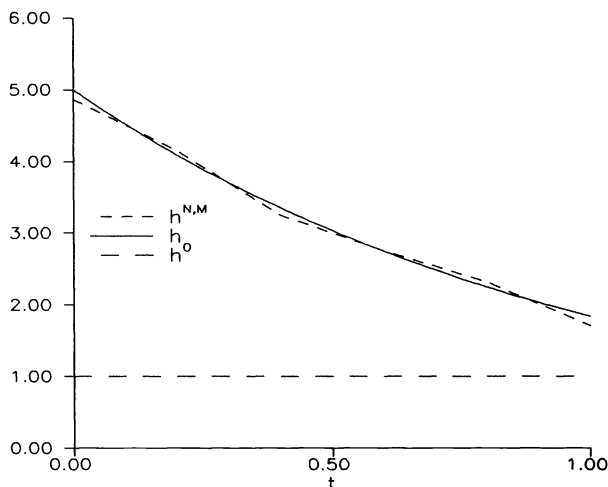


FIG. 2. Estimation of h , $M = 5$.

We use data of the form $\hat{s}_i \sim (s(t_i) - s_0)$, collected at five times from the interval $[0, 1.0]$. We approximate $h = \alpha$ over the interval $[0, 1.0]$ with $M = 5$, starting with an initial guess of $h^0 = \alpha^0 \equiv 1$. The best fit parameter we obtain is graphed in Fig. 2, along with the initial guess and true parameter. Our residual is $\tilde{J}^6(\alpha^{6.5}, h^{6.5}) = 0.114 \times 10^{-11}$.

For this example, approximately 400 function evaluations are required, with a total CP time of approximately 27 minutes. The average is 0.07 min/FE. Note that, in contrast to the biofilm (Example 1), this model equation lacks the nonlinearity ρ , and $\tilde{\mathcal{F}}$ is linear in V .

Example 3. We present a set of examples based on the model equations (3.3), which describe the diffusion of solutes through glassy polymers. In the transformed variables (fixed domain, homogeneous Dirichlet boundary condition), our version of the model equations is

$$\frac{\partial V}{\partial t} = \frac{\mathcal{D}}{s^2} \frac{\partial^2 V}{\partial y^2} + \frac{\dot{s}}{s} y \frac{\partial V}{\partial y} + \hat{F}(t, y, s, \dot{s}, \mathcal{D}, g), \quad y \in (0, 1), \quad t \in (0, T];$$

$$V(t, 0) = 0;$$

$$\frac{\mathcal{D}}{s} \frac{\partial V}{\partial y}(t, 1) + \dot{s}(t) V(t, 1) = -\dot{s}(t) \sigma(t), \quad t \in (0, T];$$

$$V(0, y) = V_0(y), \quad y \in [0, 1];$$

$$\frac{ds}{dt} = \gamma(V(t, 1)), \quad t \in [0, T];$$

$$s(0) = s_0.$$

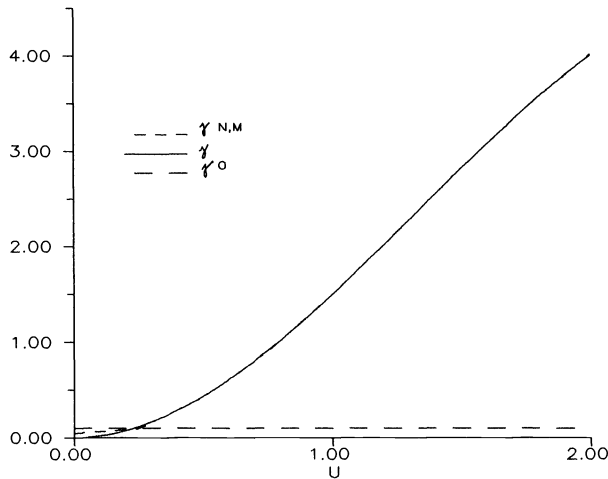
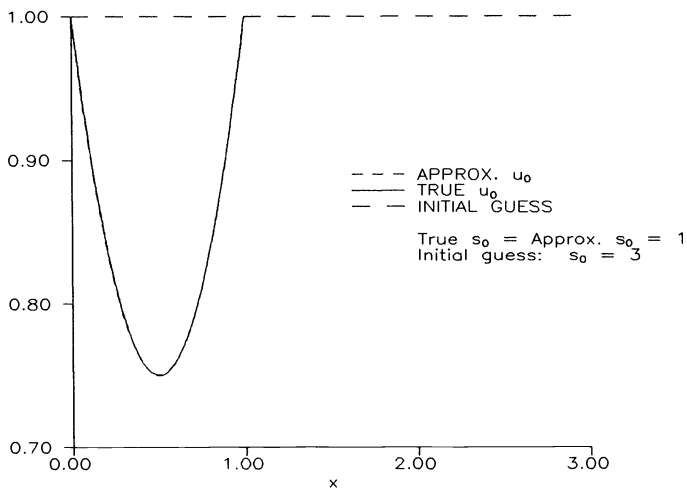
As with previous examples, for the purpose of obtaining an analytical solution we add a nonhomogeneous term F to the equation. In the two variations presented below, we estimate the nonlinear term γ and/or the initial states V_0 and s_0 .

The true γ is given by $\gamma(\theta) = 2\theta^2 - \frac{1}{2}\theta^3$, and the other parameters have been held fixed (assumed known) at $\mathcal{D} = 1$, $g \equiv 1$. In each example below, we choose the true solution U, s and then evaluate and hold fixed (assumed known) the parameter $\sigma(t) = -(1/s\dot{s})U_y(t, 1) - U(t, 1)$. As in Example 2, we construct $F(t, y) = U_t - (\mathcal{D}/s^2)U_{yy} - (\dot{s}/s)yU_y$, and then obtain

$$\hat{F} = \frac{2}{(s^N)^2} + 2 \frac{\dot{s}^N}{s^N} y(y-1) + F.$$

Example 3A. We first estimate γ without a priori parametrization. We approximate this parameter using linear splines with $M = 8$. We choose the true solution to be $U(t, y) = 2e^{-t}y + (y-1)^2$, $s(t) = \frac{1}{3}(11 - 12e^{-2t} + 4e^{-3t})$, and hold the initial conditions fixed (assumed known). Since we know the true solution, we choose $\bar{u} = 2.0$ for the interval of approximation of γ . The data we use is of the form $\hat{s}_i \sim (s(t_i) - s_0)$, collected at ten times from the interval $[0, 2.5]$. In Fig. 3(a), we graph our initial guess, the true parameter, and our estimated $\hat{\gamma}^{6,8}$. Our residual is $\tilde{J}^6 = 0.114 \times 10^{-8}$. This example takes approximately 10.5 CP minutes, with an average of 0.04 min/FE.

Example 3B. In this example, we estimate both γ and the initial conditions. To reduce the size of the parameter space, we choose an a priori parametrization for γ and approximate V_0 with linear splines using $M = 2$. We parametrize γ as $\gamma(\theta) = c_0 + c_1\theta + c_2\theta^2 + c_3\theta^3$; the true parameter is then $c = (0.0, 0.0, 2.0, -0.5)$. The true solution

FIG. 3(a). Estimation of γ , $M = 8$.FIG. 3(b). Estimation of initial state, $M = 2$.

is $U(t, y) = e^{-t}y + (y-1)^2$, and $s(t) = \frac{1}{6}(11 - 6e^{-2t} + e^{-3t})$; thus, the true U_0 is $U_0 = y + (y-1)^2$ (so that the true $V_0(y)$ is given by $V_0(y) = y$), and the true $s_0 = 1$. Our data is assumed to be of the type $\hat{s}_i \sim (s(t_i) - s_0)$ and $\hat{u}_i \sim s(t_i) \int_0^1 U(t_i, y) dy$, collected at ten times from the interval $[0, 2.5]$.

Our initial guesses for s_0 and c are 3.0 and $(1.0, 1.0, 1.0, 1.0)$, respectively. The initial guess, as well as the true and estimated u_0 (obtained via the transformation described in Remark 4.1) are graphed in Fig. 3(b). Our estimate for s_0 is $\hat{s}_0^6 = 1.0000$ and for c is $\hat{c}^6 = (0.00000, -0.00005, 2.00021, -0.50026)$. We note that the “true” and approximate u_0 are indistinguishable because the “true” V_0 (V_0 is the quantity with which we perform our estimation procedure) is, in fact, linear, so we are able to represent it exactly within the approximation space. Thus the transformation from \hat{V}_0^M to \hat{u}_0 gives us an essentially exact reconstruction. Our residual is $\tilde{J}^6 = 0.542 \times 10^{-10}$. This example requires approximately 8 CP minutes with an average of 0.01 min/FE.

Example 4. We consider a slight modification of an example of the Stefan problem, which appears in [15]. The transformed model equations are

$$\frac{\partial U}{\partial t} = \frac{1}{s^2} \frac{\partial}{\partial y} \left(\mathcal{D} \frac{\partial U}{\partial y} \right) + \frac{\dot{s}}{s} y \frac{\partial U}{\partial y}, \quad y \in (0, 1), \quad t \in (0, T];$$

$$\frac{1}{s} \frac{\partial U}{\partial y}(t, 0) = g(t);$$

$$U(t, 1) = 0; \quad t \in (0, T];$$

$$U(0, y) = U_0(y), \quad y \in [0, 1];$$

$$\frac{ds}{dt} = -\frac{\mathcal{D}}{s} \frac{\partial U}{\partial y}(t, 1), \quad t \in [0, T];$$

$$s(0) = s_0.$$

The parameters we consider unknown here are $\mathcal{D} = \mathcal{D}(t)$, and the initial location of the boundary s_0 . We approximate the time-varying coefficient with linear splines with $M = 5$. The true diffusion coefficient is $\mathcal{D}(t) = \cos t$, the parameter g is assumed known and is given by $g(t) = -\exp(1 + \sin t)$, and the true solution is given by $U(t, y) = \exp((1 - y)(1 + \sin t)) - 1$ and $s(t) = 1 + \sin t$.

Example 4A. Here we take for our data, observations of $u(t_i, x_j)$ at six times in $[0, 1.0]$ and at $x = 0.0, 0.25$, and 0.50 . We estimate both \mathcal{D} and s_0 . The initial guess, approximate and true coefficient are all plotted in Fig. 4(a). For the initial boundary location, we guess $s_0^0 = 0.3$ and obtain the estimate $\hat{s}_0 = 0.9933$ (the “true” value is 1.0). The residual is $\tilde{J}^6(\hat{\mathcal{D}}^{6,5}, \hat{s}_0) = 0.104 \times 10^{-3}$.

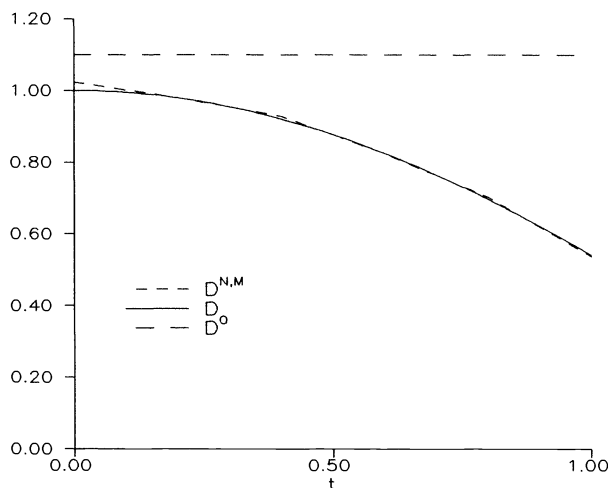
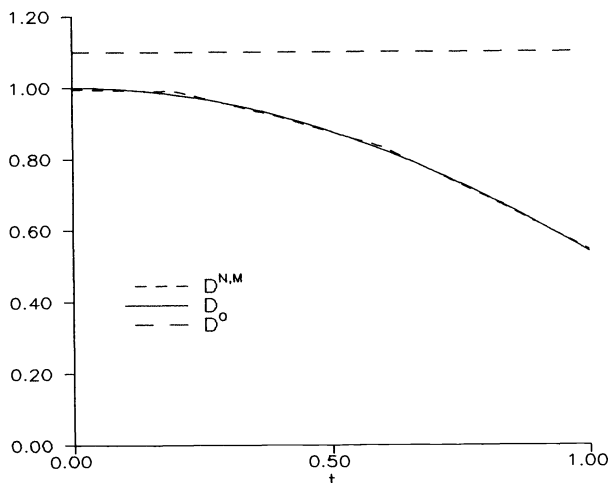
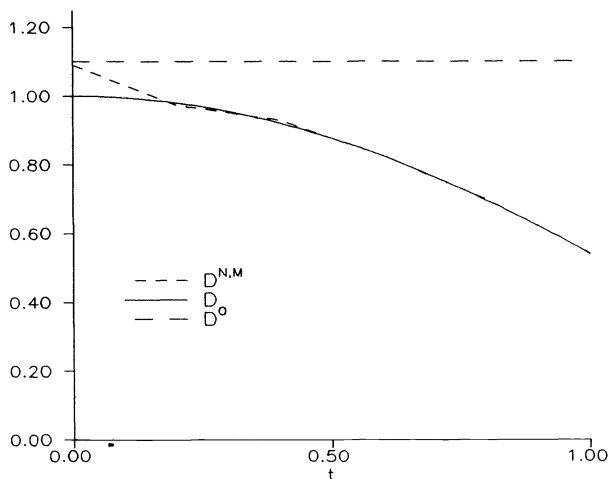


FIG. 4(a). Estimation of $D(t)$, $M = 5$.

Example 4B. Here we take the more realistic data set $u(t_i, 0)$ with six times from $[0, 1.0]$. Estimating only \mathcal{D} , assuming s_0 is known, we obtain the estimate shown in Fig. 4(b). Here our residual is $\tilde{J}^6(\hat{\mathcal{D}}^{6,5}) = 0.276 \times 10^{-6}$.

Example 4C. Finally, we use the data $u(t_i, 0)$, with ten times from $[0, 1.0]$ and estimate both \mathcal{D} and s_0 . Our initial guess, approximation, and true diffusion coefficients are plotted in Fig. 4(c), and our initial guess and estimate for s_0 are given by $s_0^0 = 0.3$, $\hat{s}_0 = 0.985$. Here our residual is $\tilde{J}^6(\hat{\mathcal{D}}^{6,5}, \hat{s}_0) = 0.500 \times 10^{-5}$.

Example 5. Here we consider the oxygen diffusion equations (3.5). We obtain our “data” from [10] (the numerical results of many authors are reported in Crank [10]).

FIG. 4(b). Estimation of $D(t)$, $M=5$.FIG. 4(c). Estimation of $D(t)$, $M=5$.

For our states, we use both cubic and quintic splines (note that linear splines are out of the question, due to the form of $\tilde{\mathcal{F}}$); quintics are superior and used in all examples reported here.

The original equations (3.5) are equivalent via the change of variables (see [10, p. 20]) $X = x/s_0$, $\tau = (D/s_0^2)t$, $w = (D/(\rho s_0^2))u \equiv 2u$, and $S = s/s_0$ to the following system:

$$\frac{\partial w}{\partial \tau} = \frac{\partial^2 w}{\partial X^2} - 1, \quad X \in [0, S(\tau)], \quad \tau \in [0, T];$$

$$\frac{\partial w}{\partial X}(\tau, 0) = 0;$$

$$w(\tau, S(\tau)) = 0, \quad \tau \in [0, T];$$

$$w(0, X) = \frac{1}{2}(1-X)^2, \quad X \in [0, 1];$$

$$\frac{dS}{d\tau} = -\frac{\partial^3 w}{\partial X^3}(\tau, S(\tau)), \quad \tau \in [0, T];$$

$$S(0) = 1.$$

Let $q = \rho$; we can see that, having only observations of u (a solution of (3.5)) at $x = 0$, at a series of times t_i , is equivalent to having observations of $\frac{1}{2}w$ (where w satisfies the system given above) at $X = 0$, at the series of times $\tau_i = (q/2)t_i$. Since the above system contains no parameters, it is clear that we could only hope to estimate the time-scaling factor q . Thus, at least for the simplified model equations of (3.5), while there are several parameters that could be unknown and of interest to estimate, it will be possible to estimate only one.

We do this successfully. In one example, we choose the true $q = 2.0$ and have data at six values of time within $[0, 0.16]$, all at the left boundary ($x = X = 0$). Beginning with the initial guess $q^0 = 0.3$, we obtain the estimate $q^6 = 1.999947$. Our residual is $\hat{J}^6 = 0.290 \times 10^{-9}$.

In several examples, we also try to estimate two of the constants D , ρ , and s_0 , while holding the third fixed. In all cases, we obtain convergence, but not to the "true" parameters; as expected, there are equivalence classes of parameters, so that, for example, having estimated $\hat{\rho}$ and \hat{s}_0 , we obtain convergence to values satisfying $\hat{\rho}\hat{s}_0^2 = \rho^*(s_0^*)^2$ (where the $*$ represents the "true" value).

8. Conclusion. For moving boundary problems within a certain class, we develop an approximation method for the purpose of estimating unknown parameters, given observations of the system. We prove that the approximations converge in a sense that guarantees that the estimates we compute by fitting the approximate model equations to the data give relevant information to the true estimation problem.

We choose a set of test examples to illustrate the efficacy of the method. We estimate variable parameters without the necessity of assuming a priori parametrizations (both space and time dependence); these parameters may appear in the partial differential equation or in the boundary conditions. We estimate unknown initial conditions and nonlinearities (i.e., state dependence), again without the necessity of a priori parametrizations. We use several types of data corresponding to various experimental observations.

In addition to numerically demonstrating the results we have proved theoretically, we present evidence that the method could be useful even in cases for which there is no theory; specifically, we use data sets that involve a stronger topology than that of our convergence result, and we successfully estimate parameters in models for the Stefan problem and oxygen diffusion. We are currently exploring a variation of the method discussed here, which may be applied to the Stefan and oxygen diffusion problems, and for which theoretical results can be proved.

Appendix. We present here the proofs of Proposition 4.1 and the theorems of § 5.

Proof of Proposition 4.1. Fix $q \in \mathcal{Q}$. Let (V, s) and (\tilde{V}, \tilde{s}) be two solutions of (4.1); i.e., for each $t \in (0, T]$ and all $\phi \in H_B^1(0, 1)$, $V(t, \cdot) \in H_B^1(0, 1)$ and s satisfy

$$\begin{aligned} \left\langle \frac{\partial V}{\partial t}, \phi \right\rangle &= -\frac{1}{s^2} \left\langle \mathcal{D}, \frac{\partial V}{\partial y}, \frac{\partial \phi}{\partial y} \right\rangle - \frac{1}{s} \left\langle \mathcal{V}_1 V, \frac{\partial \phi}{\partial y} \right\rangle + \frac{\mathcal{V}_2(s)}{s} \left\langle \frac{\partial V}{\partial y}, \phi \right\rangle + \frac{\dot{s}}{s} \left\langle y \frac{\partial V}{\partial y}, \phi \right\rangle \\ &\quad + \langle \rho(V - \tilde{F}), \phi \rangle + \langle \hat{F}(s, \tilde{s}), \phi \rangle + \frac{1}{s} (-\alpha_2 V(t, 1) + h) \phi(1), \\ \frac{ds}{dt} &= \tilde{\mathcal{F}}(s, V - \tilde{F}(g); \Gamma), \\ V(0, y) &= U_0(y) + \tilde{F}(0, y, g), \\ s(0) &= s_0, \end{aligned}$$

and $\tilde{V}(t, \cdot) \in H_B^1(0, 1)$, and \tilde{s} satisfy

$$\begin{aligned} \left\langle \frac{\partial \tilde{V}}{\partial t}, \phi \right\rangle = & -\frac{1}{\tilde{s}^2} \left\langle \mathcal{D} \frac{\partial \tilde{V}}{\partial y}, \frac{\partial \phi}{\partial y} \right\rangle - \frac{1}{\tilde{s}} \left\langle \mathcal{V}_1 \tilde{V}, \frac{\partial \phi}{\partial y} \right\rangle + \frac{\mathcal{V}_2(\tilde{s})}{\tilde{s}} \left\langle \frac{\partial \tilde{V}}{\partial y}, \phi \right\rangle + \frac{\dot{\tilde{s}}}{\tilde{s}} \left\langle y \frac{\partial \tilde{V}}{\partial y}, \phi \right\rangle \\ & + \langle \rho(\tilde{V} - \tilde{F}), \phi \rangle + \langle \tilde{F}(\tilde{s}, \dot{\tilde{s}}), \phi \rangle + \frac{1}{\tilde{s}} (-\alpha_2 \tilde{V}(t, 1) + h) \phi(1), \end{aligned}$$

$$\frac{d\tilde{s}}{dt} = \tilde{\mathcal{F}}(\tilde{s}, \tilde{V} - \tilde{F}(g); \Gamma), \quad \tilde{V}(0, y) = U_0(y) + \tilde{F}(0, y, g), \quad \tilde{s}(0) = s_0.$$

Subtracting the above equations and setting $W = V - \tilde{V}$, $r = s - \tilde{s}$, we have

$$\begin{aligned} \left\langle \frac{\partial W}{\partial t}, \phi \right\rangle = & -\frac{1}{s^2} \left\langle \mathcal{D} \frac{\partial W}{\partial y}, \frac{\partial \phi}{\partial y} \right\rangle - \frac{1}{s} \left\langle \mathcal{V}_1 W, \frac{\partial \phi}{\partial y} \right\rangle + \frac{\mathcal{V}_2(s)}{s} \left\langle \frac{\partial W}{\partial y}, \phi \right\rangle + \frac{\dot{s}}{s} \left\langle y \frac{\partial W}{\partial y}, \phi \right\rangle \\ & - \left(\frac{1}{s^2} - \frac{1}{\tilde{s}^2} \right) \left\langle \mathcal{D} \frac{\partial \tilde{V}}{\partial y}, \frac{\partial \phi}{\partial y} \right\rangle - \left(\frac{1}{s} - \frac{1}{\tilde{s}} \right) \left\langle \mathcal{V}_1 \tilde{V}, \frac{\partial \phi}{\partial y} \right\rangle \\ & + \left(\frac{\mathcal{V}_2(s)}{s} - \frac{\mathcal{V}_2(\tilde{s})}{\tilde{s}} \right) \left\langle \frac{\partial \tilde{V}}{\partial y}, \phi \right\rangle + \left(\frac{\dot{s}}{s} - \frac{\dot{\tilde{s}}}{\tilde{s}} \right) \left\langle y \frac{\partial \tilde{V}}{\partial y}, \phi \right\rangle \\ (A.1) \quad & + \rho(V - \tilde{V}) - \rho(\tilde{V} - \tilde{F}), \phi \rangle + \langle \hat{F}(s, \dot{s}) - \hat{F}(\tilde{s}, \dot{\tilde{s}}), \phi \rangle \\ & - \frac{\alpha_2}{s} W(t, 1) \phi(1) - \alpha_2 \left(\frac{1}{s} - \frac{1}{\tilde{s}} \right) \tilde{V}(t, 1) \phi(1) + h \left(\frac{1}{s} - \frac{1}{\tilde{s}} \right) \phi(1), \end{aligned}$$

$$\frac{dr}{dt} = \tilde{\mathcal{F}}(s, V - \tilde{F}(g); \Gamma) - \tilde{\mathcal{F}}(\tilde{s}, \tilde{V} - \tilde{F}(g); \Gamma), \quad W(0, y) \equiv 0, \quad r(0) = 0.$$

Now we choose $\phi = W$ and use the fact that for any constant $c > 0$ and $f, g \in H^0(0, 1)$, we have $\langle f, g \rangle \leq (1/2c) \|f\|^2 + (c/2) \|g\|^2$ to obtain

$$\begin{aligned} \frac{1}{2} \frac{d}{dt} \|W\|^2 \leq & -\frac{D_L}{s^2} \left\| \frac{\partial W}{\partial y} \right\|^2 + \frac{|\mathcal{V}_1|_\infty^2}{2cs^2} \|W\|^2 + \frac{c}{2} \left\| \frac{\partial W}{\partial y} \right\|^2 \\ & + \frac{1}{2c} \left(\frac{\mathcal{V}_2(\dot{s})}{s} \right)^2 \|W\|^2 + \frac{c}{2} \left\| \frac{\partial W}{\partial y} \right\|^2 + \frac{1}{2c} \left(\frac{\dot{s}}{s} \right)^2 \|W\|^2 + \frac{c}{2} \left\| \frac{\partial W}{\partial y} \right\|^2 \\ & + \frac{1}{2c} |\mathcal{D}|_\infty^2 \left\| \frac{\partial \tilde{V}}{\partial y} \right\|^2 \frac{(s + \tilde{s})^2}{(s\tilde{s})^4} r^2 + \frac{c}{2} \left\| \frac{\partial W}{\partial y} \right\|^2 \\ & + \frac{1}{2c} |\mathcal{V}_1|_\infty^2 \|\tilde{V}\|^2 \frac{1}{(s\tilde{s})^2} r^2 + \frac{c}{2} \left\| \frac{\partial W}{\partial y} \right\|^2 + \frac{1}{2} \left(\frac{\mathcal{V}_2(\dot{s})}{s\tilde{s}} \right)^2 \left\| \frac{\partial \tilde{V}}{\partial y} \right\|^2 r^2 \\ (A.2) \quad & + \frac{1}{2} \|W\|^2 + \frac{c}{2} \left(\frac{\mathcal{V}_2^{\text{Lip}}}{\tilde{s}} \right)^2 \left\| \frac{\partial \tilde{V}}{\partial y} \right\|^2 r^2 + \frac{1}{2c} \|W\|^2 + \frac{c}{2s^2} \left\| \frac{\partial \tilde{V}}{\partial y} \right\|^2 r^2 \\ & + \frac{1}{2c} \|W\|^2 + \frac{1}{2} \left(\frac{\dot{s}}{s\tilde{s}} \right)^2 \left\| \frac{\partial \tilde{V}}{\partial y} \right\|^2 r^2 + \frac{1}{2} \|W\|^2 + \rho^{\text{Lip}} \|W\|^2 \\ & + \frac{c}{2} \|\hat{F}(s, \dot{s}) - \hat{F}(\tilde{s}, \dot{\tilde{s}})\|^2 + \frac{1}{2c} \|W\|^2 \\ & - \frac{\alpha_2}{s} W^2(t, 1) + \frac{1}{2c} \left(\frac{\alpha_2}{s\tilde{s}} \right)^2 \tilde{V}^2(t, 1) r^2 + \frac{c}{2} W^2(t, 1) \\ & + \frac{1}{2c} \left(\frac{h}{s\tilde{s}} \right)^2 r^2 + \frac{c}{2} W^2(t, 1). \end{aligned}$$

Using (H \mathcal{F}), we bound \dot{r} by $|\dot{r}| \leq \mu(|V - \tilde{F}|_\infty, \Gamma)|r| + \lambda(|\tilde{s}|, \Gamma)|W|_\infty$, from which it follows that

$$\begin{aligned} \frac{1}{2} \frac{d}{dt}(r^2) &\leq \mu(|V - \tilde{F}|_\infty, \Gamma)r^2 + \frac{\lambda^2(|\tilde{s}|, \Gamma)}{2c} r^2 + \frac{c}{2} |W|_\infty^2 \\ &\leq \mu(|V - \tilde{F}|_\infty, \Gamma)r^2 + \frac{\lambda^2(|\tilde{s}|, \Gamma)}{2c} r^2 + c \left\| \frac{\partial W}{\partial y} \right\|^2 + c \|W\|^2. \end{aligned}$$

Finally, it is easily seen that there exist $\mathcal{C}_i \in L^2(0, T)$ such that

$$\begin{aligned} \|\hat{F}(s, \dot{s}) - \hat{F}(\tilde{s}, \dot{\tilde{s}})\|^2 &\leq \mathcal{C}_1 \left(\frac{1}{s^2} - \frac{1}{\tilde{s}^2} \right)^2 + \mathcal{C}_2 \left(\frac{1}{s} - \frac{1}{\tilde{s}} \right)^2 + \mathcal{C}_3 \left(\frac{\mathcal{V}_2(\dot{s})}{s} - \frac{\mathcal{V}_2(\dot{\tilde{s}})}{\tilde{s}} \right)^2 + \mathcal{C}_4 \left(\frac{\dot{s}}{s} - \frac{\dot{\tilde{s}}}{\tilde{s}} \right)^2 \\ &\leq \frac{\mathcal{C}_1(s + \tilde{s})^2}{(s\tilde{s})^4} r^2 + \frac{\mathcal{C}_2}{(s\tilde{s})^2} r^2 + 2\mathcal{C}_3 \left(\frac{\mathcal{V}_2(\dot{s})}{s\tilde{s}} \right)^2 r^2 + 2\mathcal{C}_3 \left(\frac{\mathcal{V}_2^{\text{Lip}}}{\tilde{s}} \right)^2 \dot{r}^2 \\ &\quad + \frac{2\mathcal{C}_4}{s^2} \dot{r}^2 + 2\mathcal{C}_4 \left(\frac{\dot{\tilde{s}}}{s\tilde{s}} \right)^2 r^2. \end{aligned}$$

Combining all of the above estimates, we obtain

$$\begin{aligned} \frac{1}{2} \frac{d}{dt}(\|W\|^2 + r^2) &\leq \left(-\frac{D_L}{s^2} + \frac{7}{2}c \right) \left\| \frac{\partial W}{\partial y} \right\|^2 + \left(-\frac{\alpha_2}{s} + c \right) W^2(t, 1) \\ &\quad + \left(\frac{|\mathcal{V}_1|_\infty^2}{2cs^2} + \frac{1}{2c} \left(\frac{\mathcal{V}_2(\dot{s})}{s} \right)^2 + \frac{1}{2c} \left(\frac{\dot{s}}{s} \right)^2 + \rho^{\text{Lip}} + \frac{3}{2c} + 1 + c \right) \|W\|^2 \\ &\quad + \left(\frac{1}{2c} |\mathcal{D}|_\infty^2 \left\| \frac{\partial \tilde{V}}{\partial y} \right\|^2 \frac{(s + \tilde{s})^2}{(s\tilde{s})^4} + \frac{1}{2c} |\mathcal{V}_1|_\infty^2 \|\tilde{V}\|^2 \frac{1}{(s\tilde{s})^2} \right. \\ &\quad \left. + \frac{1}{2} \left(\frac{\mathcal{V}_2(\dot{s})}{s\tilde{s}} \right)^2 \left\| \frac{\partial \tilde{V}}{\partial y} \right\|^2 + \frac{1}{2} \left(\frac{\dot{s}}{s\tilde{s}} \right)^2 \left\| \frac{\partial \tilde{V}}{\partial y} \right\|^2 + \frac{c\mathcal{C}_1(s + \tilde{s})^2}{2(s\tilde{s})^4} + \frac{c\mathcal{C}_2}{2(s\tilde{s})^2} + \mu \right. \\ &\quad \left. + \frac{\lambda^2}{2c} + c\mathcal{C}_3 \left(\frac{\mathcal{V}_2(\dot{s})}{s\tilde{s}} \right)^2 + c\mathcal{C}_4 \left(\frac{\dot{s}}{s\tilde{s}} \right)^2 + \frac{1}{2c} \left(\frac{\alpha_2}{s\tilde{s}} \right)^2 \tilde{V}^2(t, 1) + \frac{1}{2c} \left(\frac{h}{s\tilde{s}} \right)^2 \right) r^2 \\ &\quad + \left(\frac{c}{2} \left(\frac{\mathcal{V}_2^{\text{Lip}}}{\tilde{s}} \right)^2 \left\| \frac{\partial \tilde{V}}{\partial y} \right\|^2 + \frac{c}{2s^2} \left\| \frac{\partial \tilde{V}}{\partial y} \right\|^2 + c\mathcal{C}_3 \left(\frac{\mathcal{V}_2^{\text{Lip}}}{\tilde{s}} \right)^2 + \frac{c\mathcal{C}_4}{s^2} \right) \dot{r}^2 \\ &\leq \left(-\frac{D_L}{s^2} + \frac{7}{2}c \right) \left\| \frac{\partial W}{\partial y} \right\|^2 + \left(-\frac{\alpha_2}{s} + c \right) W^2(t, 1) \\ &\quad + \left(\frac{|\mathcal{V}_1|_\infty^2}{2cs^2} + \frac{1}{2c} \left(\frac{\mathcal{V}_2(\dot{s})}{s} \right)^2 + \frac{1}{2c} \left(\frac{\dot{s}}{s} \right)^2 + \rho^{\text{Lip}} + \frac{3}{2c} + 1 + c \right) \|W\|^2 \\ &\quad + \left(\frac{1}{2c} |\mathcal{D}|_\infty^2 \left\| \frac{\partial \tilde{V}}{\partial y} \right\|^2 \frac{(s + \tilde{s})^2}{(s\tilde{s})^4} + \frac{1}{2c} |\mathcal{V}_1|_\infty^2 \|\tilde{V}\|^2 \frac{1}{(s\tilde{s})^2} \right. \\ &\quad \left. + \frac{1}{2} \left(\frac{\mathcal{V}_2(\dot{s})}{s\tilde{s}} \right)^2 \left\| \frac{\partial \tilde{V}}{\partial y} \right\|^2 + \frac{1}{2} \left(\frac{\dot{s}}{s\tilde{s}} \right)^2 \left\| \frac{\partial \tilde{V}}{\partial y} \right\|^2 + \frac{c\mathcal{C}_1(s + \tilde{s})^2}{2(s\tilde{s})^4} + \frac{c\mathcal{C}_2}{2(s\tilde{s})^2} + \mu \right. \\ &\quad \left. + \frac{\lambda^2}{2c} + c\mathcal{C}_3 \left(\frac{\mathcal{V}_2(\dot{s})}{s\tilde{s}} \right)^2 + c\mathcal{C}_4 \left(\frac{\dot{s}}{s\tilde{s}} \right)^2 + \frac{1}{2c} \left(\frac{\alpha_2}{s\tilde{s}} \right)^2 \tilde{V}^2(t, 1) + \frac{1}{2c} \left(\frac{h}{s\tilde{s}} \right)^2 \right) r^2 \\ &\quad + c \left(\frac{1}{2} \left\{ \left(\frac{\mathcal{V}_2^{\text{Lip}}}{\tilde{s}} \right)^2 + \frac{1}{s^2} \right\} \left\| \frac{\partial \tilde{V}}{\partial y} \right\|^2 + \mathcal{C}_3 \left(\frac{\mathcal{V}_2^{\text{Lip}}}{\tilde{s}} \right)^2 + \frac{\mathcal{C}_4}{s^2} \right) \\ &\quad \cdot \left(2\mu^2 r^2 + 4\lambda^2 \|W\|^2 + 4\lambda^2 \left\| \frac{\partial W}{\partial y} \right\|^2 \right). \end{aligned}$$

Now we choose c such that the coefficients in front of $\|(\partial W)/\partial y\|^2$ and $W^2(t, 1)$ are nonpositive, and therefore conclude that there exists a function $\omega \in L^1(0, T)$ such that

$$\frac{d}{dt}(\|W\|^2 + r^2) \leq \omega(\|W\|^2 + r^2) \quad \text{for a.e. } t \in (0, T].$$

Since we have zero initial conditions for W and r , the Gronwall inequality implies that $W \equiv 0$, $r \equiv 0$, and this proves the uniqueness.

The polymer example (Example 3.3) must be treated slightly differently; here we have $h(t) = -\sigma(t)\dot{s}(t)$ and $\alpha_2(t) = \dot{s}(t)$. As with the convergence arguments of § 5, we must make use of the specific form of $\tilde{\mathcal{F}}$. As described in the paragraph following Theorem 5.3, we rely on the (physically reasonable) assumptions that γ is positive and nondecreasing and that the solution satisfies $V(t, 1) \geq 0$.

We begin the uniqueness proof in the same way; however, the last line of equation (A.1) is now

$$-\frac{\dot{s}}{s} W(t, 1)\phi(1) - \left(\frac{\dot{s}}{s} - \frac{\dot{\tilde{s}}}{\tilde{s}}\right) \tilde{V}(t, 1)\phi(1) - \sigma\left(\frac{\dot{s}}{s} - \frac{\dot{\tilde{s}}}{\tilde{s}}\right)\phi(1),$$

which, with the choice $\phi = W$, becomes

$$\begin{aligned} & -\frac{\dot{s}}{s} W^2(t, 1) - \left(\frac{\dot{s}}{s} - \frac{\dot{\tilde{s}}}{\tilde{s}}\right) (\tilde{V}(t, 1) + \sigma) W(t, 1) \\ &= -\frac{\dot{s}}{s} W^2(t, 1) + \frac{\dot{s}}{s\tilde{s}} (\tilde{V}(t, 1) + \sigma) W(t, 1)r - \frac{1}{\tilde{s}} (\dot{s} - \dot{\tilde{s}}) (\tilde{V}(t, 1) + \sigma) W(t, 1) \\ &\leq -\frac{\dot{s}}{s} W^2(t, 1) + \frac{1}{2c} \left(\frac{\dot{s}(\tilde{V}(t, 1) + \sigma)}{s\tilde{s}}\right)^2 r^2 + \frac{c}{2} W^2(t, 1) \\ &\quad - \frac{1}{\tilde{s}} (\tilde{V}(t, 1) + \sigma) (\gamma(V(t, 1)) - \gamma(\tilde{V}(t, 1))) W(t, 1). \end{aligned}$$

Because γ is nondecreasing, we see that

$$(\gamma(V(t, 1)) - \gamma(\tilde{V}(t, 1))) W(t, 1) = (\gamma(V(t, 1)) - \gamma(\tilde{V}(t, 1)))(V(t, 1) - \tilde{V}(t, 1)) \geq 0.$$

Moreover, $\tilde{V} \geq 0$, $\sigma > 0$, $\tilde{s} > 0$, implies that

$$\begin{aligned} & -\frac{\dot{s}}{s} W^2(t, 1) - \left(\frac{\dot{s}}{s} - \frac{\dot{\tilde{s}}}{\tilde{s}}\right) (\tilde{V}(t, 1) + \sigma) W(t, 1) \\ &\leq -\frac{\dot{s}}{s} W^2(t, 1) + \frac{1}{2c} \left(\frac{\dot{s}(\tilde{V}(t, 1) + \sigma)}{s\tilde{s}}\right)^2 r^2 + \frac{c}{2} W^2(t, 1). \end{aligned}$$

We replace the last two lines of inequality (A.2) by the right-hand side of the above inequality. Then the coefficients change, but the arguments are the same. \square

To make the proof of Theorem 5.3 somewhat shorter and clearer, we prove some of the necessary estimates in the form of lemmas. In the first lemma we obtain a bound on the nonhomogeneous term (arising from the change of variables used to obtain homogeneous Dirichlet boundary conditions) in terms of parameter differences and states.

LEMMA A.1. *Given $\tilde{s} \in C^1[0, T]$ satisfying (HN1) and (HN2), $\phi \in H_b^1(0, 1)$, and $\tilde{q}, q \in \mathcal{Q}$, and given any positive constant c , it is possible to determine a set of constants C_i and a set of functions m_i in $H^0(0, T)$, which depend on \underline{s} , \bar{s} , K_s , $\mathcal{V}_2^{\text{Lip}}$, \tilde{q} , and c , but*

not on q , such that

$$\begin{aligned} \langle \hat{F}(\tilde{s}, \tilde{q}) - \hat{F}(s, q), \phi \rangle \leq & m_1 |\tilde{g} - g|^2 + C_1 |\dot{\tilde{g}} - \dot{g}|^2 + C_2 g^2 |\tilde{\mathcal{D}} - \mathcal{D}|_\infty^2 + C_3 g^2 |\tilde{\mathcal{V}}_1 - \mathcal{V}_1|_\infty^2 \\ & + C_4 g^2 |\tilde{\mathcal{V}}_2(\dot{\tilde{s}}) - \mathcal{V}_2(\dot{s})|^2 + C_5 \|\phi\|^2 \\ & + (m_2 + m_3 g^2 + C_6 g^2 (\mathcal{V}_2(\dot{s}))^2) |\tilde{s} - s|^2 \\ & + 3c \left\| \frac{\partial \phi}{\partial y} \right\|^2 + C_7 g^2 c |\dot{\tilde{s}} - \dot{s}|^2 + 2c g^2 \left| \frac{\dot{\tilde{s}}}{\tilde{s}} - \frac{\dot{s}}{s} \right|^2. \end{aligned}$$

Proof. Beginning with the definitions of \hat{F} and \tilde{F} (see (4.2)), using an integration by parts and the fact that by construction $\tilde{F} \in H_B^1$, we calculate

$$\begin{aligned} \langle \hat{F}(\tilde{s}, \tilde{q}) - \hat{F}(s, q), \phi \rangle = & \{(\dot{\tilde{g}} - \dot{g})(\cdot - 1)^2, \phi\} \\ & - \left\{ \frac{2\tilde{g}}{\tilde{s}^2} \left\langle \tilde{\mathcal{D}}(\cdot - 1), \frac{\partial \phi}{\partial y} \right\rangle - \frac{2g}{s^2} \left\langle \mathcal{D}(\cdot - 1), \frac{\partial \phi}{\partial y} \right\rangle \right\} \\ & - \left\{ \frac{\tilde{g}}{\tilde{s}} \left\langle \tilde{\mathcal{V}}_1(\cdot - 1)^2, \frac{\partial \phi}{\partial y} \right\rangle - \frac{g}{s} \left\langle \mathcal{V}_1(\cdot - 1)^2, \frac{\partial \phi}{\partial y} \right\rangle \right\} \\ & + \left\{ 2 \left(\frac{\tilde{g}\tilde{\mathcal{V}}_2}{\tilde{s}} - \frac{g\mathcal{V}_2}{s} \right) (\cdot - 1), \phi \right\} \\ & + \left\{ 2 \left(\tilde{g} \frac{\dot{\tilde{s}}}{\tilde{s}} - g \frac{\dot{s}}{s} \right) (\cdot - 1), \phi \right\}. \end{aligned} \quad (\text{A.3})$$

The first bracketed quantity can be bound as $(\dot{\tilde{g}} - \dot{g})(\cdot - 1)^2, \phi \rangle \leq \frac{1}{2} |\dot{\tilde{g}} - \dot{g}|^2 + \frac{1}{2} \|\phi\|^2$. Consider the second bracketed quantity

$$\begin{aligned} & \frac{2\tilde{g}}{\tilde{s}^2} \left\langle \tilde{\mathcal{D}}(\cdot - 1), \frac{\partial \phi}{\partial y} \right\rangle - \frac{2g}{s^2} \left\langle \mathcal{D}(\cdot - 1), \frac{\partial \phi}{\partial y} \right\rangle \\ & = 2 \left(\frac{\tilde{g}}{\tilde{s}^2} - \frac{g}{s^2} \right) \left\langle \tilde{\mathcal{D}}(\cdot - 1), \frac{\partial \phi}{\partial y} \right\rangle + \frac{2g}{s^2} \left\langle (\tilde{\mathcal{D}} - \mathcal{D})(\cdot - 1), \frac{\partial \phi}{\partial y} \right\rangle. \end{aligned}$$

Using the fact that for any constant $c > 0$ and $f, g \in H^0(0, 1)$, it is the case that $\langle f, g \rangle \leq (1/2c) \|f\|^2 + (c/2) \|g\|^2$, we can then bound this term as

$$\begin{aligned} & \frac{2\tilde{g}}{\tilde{s}^2} \left\langle \tilde{\mathcal{D}}(\cdot - 1), \frac{\partial \phi}{\partial y} \right\rangle - \frac{2g}{s^2} \left\langle \mathcal{D}(\cdot - 1), \frac{\partial \phi}{\partial y} \right\rangle \\ & \leq \frac{|\tilde{\mathcal{D}}|_\infty^2}{c} \left| \frac{\tilde{g}}{\tilde{s}^2} - \frac{g}{s^2} \right|^2 + c \left\| \frac{\partial \phi}{\partial y} \right\|^2 + \frac{g^2}{cs^4} |\tilde{\mathcal{D}} - \mathcal{D}|_\infty^2 + c \left\| \frac{\partial \phi}{\partial y} \right\|^2. \end{aligned}$$

With further manipulations and applying triangle inequalities, we obtain

$$\begin{aligned} & \frac{2\tilde{g}}{\tilde{s}^2} \left\langle \tilde{\mathcal{D}}(\cdot - 1), \frac{\partial \phi}{\partial y} \right\rangle - \frac{2g}{s^2} \left\langle \mathcal{D}(\cdot - 1), \frac{\partial \phi}{\partial y} \right\rangle \\ & \leq \frac{|\tilde{\mathcal{D}}|_\infty^2}{c} \frac{1}{(\tilde{s}s)^2} |\tilde{g}(s^2 - \tilde{s}^2) + (\tilde{g} - g)\tilde{s}^2|^2 + \frac{g^2}{cs^4} |\tilde{\mathcal{D}} - \mathcal{D}|_\infty^2 + 2c \left\| \frac{\partial \phi}{\partial y} \right\|^2 \\ & \leq \frac{|\tilde{\mathcal{D}}|_\infty^2}{cs^4} (2|\tilde{g}|^2 |s^2 - \tilde{s}^2|^2 + 2\tilde{s}^4 |\tilde{g} - g|^2) + \frac{g^2}{cs^4} |\tilde{\mathcal{D}} - \mathcal{D}|_\infty^2 + 2c \left\| \frac{\partial \phi}{\partial y} \right\|^2 \\ & \leq \frac{2|\tilde{\mathcal{D}}|_\infty^2}{cs^4} (|\tilde{g}|^2 4\tilde{s}^2 |s - \tilde{s}|^2 + \tilde{s}^4 |\tilde{g} - g|^2) + \frac{g^2}{cs^4} |\tilde{\mathcal{D}} - \mathcal{D}|_\infty^2 + 2c \left\| \frac{\partial \phi}{\partial y} \right\|^2. \end{aligned}$$

In an analogous manner, we can bound the third bracketed term in (A.3) as follows:

$$\begin{aligned} & \frac{\tilde{g}}{\tilde{s}} \left\langle \tilde{\mathcal{V}}_1(\cdot - 1)^2, \frac{\partial \phi}{\partial y} \right\rangle - \frac{g}{s} \left\langle \mathcal{V}_1(\cdot - 1)^2, \frac{\partial \phi}{\partial y} \right\rangle \\ & \leq \frac{|\tilde{\mathcal{V}}_1|_\infty^2}{cs^2} \left(|\tilde{g} - g|^2 + \frac{g^2}{\tilde{s}^2} |s - \tilde{s}|^2 \right) + c \left\| \frac{\partial \phi}{\partial y} \right\|^2 + \frac{g^2}{2cs^2} |\tilde{\mathcal{V}}_1 - \mathcal{V}_1|_\infty^2. \end{aligned}$$

We now obtain the following bound for the fourth bracketed term in (A.3):

$$\begin{aligned}
 2\left(\frac{\tilde{g}\tilde{\mathcal{V}}_2}{\tilde{s}} - \frac{g\mathcal{V}_2}{s}\right)\langle(\cdot-1), \phi\rangle &\leq c\left|\frac{1}{\tilde{s}}(\tilde{g}\tilde{\mathcal{V}}_2 - g\mathcal{V}_2) + \frac{g\mathcal{V}_2}{s\tilde{s}}(s-\tilde{s})\right|^2 + \frac{1}{c}\|\phi\|^2 \\
 &\leq \frac{2cg^2(\mathcal{V}_2(\dot{s}))^2}{\underline{s}^4}|s-\tilde{s}|^2 + \frac{1}{c}\|\phi\|^2 \\
 &\quad + \frac{2c}{\underline{s}^2}|(\tilde{g}-g)\tilde{\mathcal{V}}_2(\dot{s}) + g(\tilde{\mathcal{V}}_2(\dot{s}) - \mathcal{V}_2(\dot{s})) + g(\mathcal{V}_2(\dot{s}) - \mathcal{V}_2(\dot{s}))|^2 \\
 &\leq \frac{2cg^2(\mathcal{V}_2(\dot{s}))^2}{\underline{s}^4}|s-\tilde{s}|^2 + \frac{1}{c}\|\phi\|^2 \\
 &\quad + \frac{4c(\tilde{\mathcal{V}}_2(\dot{s}))^2}{\underline{s}^2}|\tilde{g}-g|^2 + \frac{8cg^2}{\underline{s}^2}|\mathcal{V}_2(\dot{s}) - \tilde{\mathcal{V}}_2(\dot{s})|^2 \\
 &\quad + \frac{8cg^2(\mathcal{V}_2^{\text{Lip}})^2}{\underline{s}^2}|\dot{s}-\dot{\tilde{s}}|^2.
 \end{aligned}$$

Recall that, by assumption, $\tilde{\mathcal{V}}_2$ is continuous in its argument, and \dot{s} is bounded. Therefore we can bound the quantity $(\tilde{\mathcal{V}}_2(\dot{s}))^2$ by a constant. Finally, it is easy to obtain the following bound on the last quantity in (A.3):

$$2\left(\tilde{g}\frac{\dot{s}}{\tilde{s}} - g\frac{\dot{s}}{s}\right)\langle(\cdot-1), \phi\rangle \leq \frac{2cK_s^2}{\underline{s}^2}|\tilde{g}-g|^2 + 2cg^2\left|\frac{\dot{s}}{\tilde{s}} - \frac{\dot{s}}{s}\right|^2 + \frac{1}{c}\|\phi\|^2.$$

By combining the above estimates, we obtain the stated result. \square

Our second lemma gives us a bound on differences in the functional describing the dynamics of the moving boundary. The proof of this lemma is a straightforward application of the triangle inequality and the properties of $\tilde{\mathcal{F}}$ and is therefore omitted.

LEMMA A.2. *Given $\tilde{s}, s \in C^1[0, T]$ satisfying (HN1) and (HN2), $\tilde{V}, V \in L^\infty(0, 1)$, $\tilde{\Gamma}, \Gamma \in \mathcal{G}$, and $\tilde{\mathcal{F}}$ satisfying (H \mathcal{F}), it follows that*

$$\begin{aligned}
 |\tilde{\mathcal{F}}(\tilde{s}, \tilde{V}, \tilde{\Gamma}) - \tilde{\mathcal{F}}(s, V, \Gamma)|^2 &\leq 2|\tilde{\mathcal{F}}(\tilde{s}, \tilde{V}, \tilde{\Gamma}) - \tilde{\mathcal{F}}(\tilde{s}, \tilde{V}, \Gamma)|^2 + 4\mu^2(|\tilde{V}|_\infty, |\Gamma|_\gamma)|\tilde{s}-s|^2 \\
 &\quad + 4\lambda^2(s, |\Gamma|_\gamma)|\tilde{V}-V|_\infty^2
 \end{aligned}$$

and

$$\begin{aligned}
 \left(\frac{1}{\tilde{s}}\tilde{\mathcal{F}}(\tilde{s}, \tilde{V}, \tilde{\Gamma}) - \frac{1}{s}\tilde{\mathcal{F}}(s, V, \Gamma)\right)^2 &\leq \frac{2}{\underline{s}^2}\left\{|\tilde{\mathcal{F}}(\tilde{s}, \tilde{V}, \tilde{\Gamma}) - \tilde{\mathcal{F}}(\tilde{s}, \tilde{V}, \Gamma)|^2 + \left(4\mu^2(|\tilde{V}|_\infty, |\Gamma|_\gamma)\right.\right. \\
 &\quad \left.\left.+ \left(\frac{K_s}{\underline{s}}\right)^2\right)|\tilde{s}-s|^2 + 4\lambda^2(s, |\Gamma|_\gamma)|\tilde{V}-V|_\infty^2\right\}.
 \end{aligned}$$

Proof of Theorem 5.3. As our hypotheses (HE1) and (HA1) guarantee that $\|P^N V - V\|^2 \rightarrow 0$, as $N \rightarrow \infty$, it is sufficient to prove that $\|P^N V - V^{N,M}\|^2 + |s - s^{N,M}|^2 \rightarrow 0$, as $N, M \rightarrow \infty$. Our proof of this convergence statement involves use of the Gronwall inequality. Our goal is to obtain a bound on $(d/dt)(\|P^N V - V^{N,M}\|^2 + |s - s^{N,M}|^2)$ in terms of $(\|P^N V - V^{N,M}\|^2 + |s - s^{N,M}|^2)$ and various quantities, which, due to our assumptions about parameter convergence and about the approximations, converge to zero in an appropriate way. We begin with the observation that

$$\left\langle \frac{\partial}{\partial t}(P^N V - V^{N,M}), \phi \right\rangle = \left\langle P^N \frac{\partial V}{\partial t}, \phi \right\rangle - \left\langle \frac{\partial V^{N,M}}{\partial t}, \phi \right\rangle.$$

Since, for any $\phi \in H^N$, $\langle P^N(\partial V/\partial t), \phi \rangle = \langle \partial V/\partial t, \phi \rangle$, we can use (4.1) and (4.1^N) to

write the following.

For $\phi \in H^N$,

$$\begin{aligned}
 \left\langle \frac{\partial}{\partial t} (P^N V - V^{N,M}), \phi \right\rangle &= -\frac{1}{s^2} \left\langle (\tilde{\mathcal{D}} - \mathcal{D}^M) \frac{\partial V}{\partial y}, \frac{\partial \phi}{\partial y} \right\rangle - \left(\frac{1}{s^2} - \frac{1}{(s^{N,M})^2} \right) \left\langle \mathcal{D}^M \frac{\partial V}{\partial y}, \frac{\partial \phi}{\partial y} \right\rangle \\
 &\quad + \frac{1}{(s^{N,M})^2} \left\langle \mathcal{D}^M \frac{\partial}{\partial y} (V^{N,M} - P^N V), \frac{\partial \phi}{\partial y} \right\rangle \\
 &\quad + \frac{1}{(s^{N,M})^2} \left\langle \mathcal{D}^M \frac{\partial}{\partial y} (P^N V - V), \frac{\partial \phi}{\partial y} \right\rangle \\
 &\quad - \frac{1}{s} \left\langle (\tilde{\mathcal{V}}_1 - \mathcal{V}_1^M) V, \frac{\partial \phi}{\partial y} \right\rangle - \left(\frac{1}{s} - \frac{1}{s^{N,M}} \right) \left\langle \mathcal{V}_1^M V, \frac{\partial \phi}{\partial y} \right\rangle \\
 &\quad + \frac{1}{s^{N,M}} \left\langle \mathcal{V}_1^M (V^{N,M} - P^N V), \frac{\partial \phi}{\partial y} \right\rangle \\
 &\quad + \frac{1}{s^{N,M}} \left\langle \mathcal{V}_1^M (P^N V - V), \frac{\partial \phi}{\partial y} \right\rangle \\
 &\quad + \frac{1}{s} (\tilde{\mathcal{V}}_2(s) - \mathcal{V}_2^M(s)) \left\langle \frac{\partial V}{\partial y}, \phi \right\rangle + \frac{1}{s} (\mathcal{V}_2^M(s) \\
 &\quad - \mathcal{V}_2^M(s^{N,M})) \left\langle \frac{\partial V}{\partial y}, \phi \right\rangle \\
 &\quad + \left(\frac{1}{s} - \frac{1}{s^{N,M}} \right) \mathcal{V}_2^M(s^{N,M}) \left\langle \frac{\partial V}{\partial y}, \phi \right\rangle \\
 &\quad + \frac{1}{s^{N,M}} \mathcal{V}_2^M(s^{N,M}) \left\langle \frac{\partial}{\partial y} (V - P^N V), \phi \right\rangle \\
 &\quad + \frac{1}{s^{N,M}} \mathcal{V}_2^M(s^{N,M}) \left\langle \frac{\partial}{\partial y} (P^N V - V^{N,M}), \phi \right\rangle \\
 &\quad + \left(\frac{\dot{s}}{s} - \frac{\dot{s}^{N,M}}{s^{N,M}} \right) \left\langle y \frac{\partial V}{\partial y}, \phi \right\rangle + \frac{\dot{s}^{N,M}}{s^{N,M}} \left\langle y \frac{\partial}{\partial y} (V - P^N V), \phi \right\rangle \\
 &\quad + \frac{\dot{s}^{N,M}}{s^{N,M}} \left\langle y \frac{\partial}{\partial y} (P^N V - V^{N,M}), \phi \right\rangle \\
 &\quad + \langle \tilde{\rho}(V - \tilde{F}(\tilde{g})) - \rho^M(V - \tilde{F}(\tilde{g})), \phi \rangle \\
 &\quad + \langle \rho^M(V - \tilde{F}(\tilde{g})) - \rho^M(P^N V - \tilde{F}(\tilde{g})), \phi \rangle \\
 &\quad + \langle \rho^M(P^N V - \tilde{F}(\tilde{g})) - \rho^M(V^{N,M} - \tilde{F}(\tilde{g})), \phi \rangle \\
 &\quad + \langle \rho^M(V^{N,M} - \tilde{F}(\tilde{g})) - \rho^M(V^{N,M} - \tilde{F}(g^M)), \phi \rangle \\
 &\quad + \langle \hat{F}(s, \tilde{q}) - \hat{F}(s^{N,M}, q^M), \phi \rangle \\
 &\quad + \frac{1}{s} (\alpha_2^M - \tilde{\alpha}_2) V(t, 1) \phi(1) + \left(\frac{1}{s^{N,M}} - \frac{1}{s} \right) \alpha_2^M V(t, 1) \phi(1) \\
 &\quad - \frac{\alpha_2^M}{s^{N,M}} (V(t, 1) - P^N V(t, 1)) \phi(1) \\
 &\quad - \frac{\alpha_2^M}{s^{N,M}} (P^N V(t, 1) - V^{N,M}(t, 1)) \phi(1) \\
 &\quad + \frac{1}{s} (\tilde{h} - h^M) \phi(1) + \left(\frac{1}{s} - \frac{1}{s^{N,M}} \right) h^M \phi(1).
 \end{aligned}$$

Choosing $\phi \equiv P^N V(t, \cdot) - V^{N,M}(t, \cdot) \in H^N$ in the above, and rewriting slightly, we can see that for each $t \in [0, T]$,

$$\begin{aligned}
 & \frac{1}{2} \frac{d}{dt} (\|\phi\|^2 + (s - s^{N,M})^2) = \left\langle \phi, \frac{\partial \phi}{\partial t} \right\rangle + (s - s^{N,M})(\dot{s} - \dot{s}^{N,M}) \\
 &= -\frac{1}{s^2} \left\langle (\tilde{\mathcal{D}} - \mathcal{D}^M) \frac{\partial V}{\partial y}, \frac{\partial \phi}{\partial y} \right\rangle - \frac{1}{(ss^{N,M})^2} ((s^{N,M})^2 - s^2) \left\langle \mathcal{D}^M \frac{\partial V}{\partial y}, \frac{\partial \phi}{\partial y} \right\rangle \\
 & \quad - \frac{1}{(s^{N,M})^2} \left\langle \mathcal{D}^M \frac{\partial \phi}{\partial y}, \frac{\partial \phi}{\partial y} \right\rangle \\
 & \quad + \frac{1}{(s^{N,M})^2} \left\langle \mathcal{D}^M \frac{\partial}{\partial y} (P^N V - V), \frac{\partial \phi}{\partial y} \right\rangle \\
 & \quad - \frac{1}{s} \left\langle (\tilde{\mathcal{V}}_1 - \mathcal{V}_1^M) V, \frac{\partial \phi}{\partial y} \right\rangle - \frac{1}{ss^{N,M}} (s^{N,M} - s) \left\langle \mathcal{V}_1^M V, \frac{\partial \phi}{\partial y} \right\rangle - \frac{1}{s^{N,M}} \left\langle \mathcal{V}_1^M \phi, \frac{\partial \phi}{\partial y} \right\rangle \\
 & \quad + \frac{1}{s^{N,M}} \left\langle \mathcal{V}_1^M (P^N V - V), \frac{\partial \phi}{\partial y} \right\rangle \\
 & \quad + \frac{1}{s} (\tilde{\mathcal{V}}_2(\dot{s}) - \mathcal{V}_2^M(\dot{s})) \left\langle \frac{\partial V}{\partial y}, \phi \right\rangle + \frac{1}{s} (\mathcal{V}_2^M(\dot{s}) - \mathcal{V}_2^M(\dot{s}^{N,M})) \left\langle \frac{\partial V}{\partial y}, \phi \right\rangle \\
 & \quad + \frac{1}{ss^{N,M}} (s^{N,M} - s) \mathcal{V}_2^M(\dot{s}^{N,M}) \left\langle \frac{\partial V}{\partial y}, \phi \right\rangle + \frac{1}{s^{N,M}} \mathcal{V}_2^M(\dot{s}^{N,M}) \\
 & \quad \cdot \left\langle \frac{\partial}{\partial y} (V - P^N V), \phi \right\rangle \\
 & \quad + \frac{1}{s^{N,M}} \mathcal{V}_2^M(\dot{s}^{N,M}) \left\langle \frac{\partial \phi}{\partial y}, \phi \right\rangle \\
 & \quad + \left(\frac{\dot{s}}{s} - \frac{\dot{s}^{N,M}}{s^{N,M}} \right) \left\langle y \frac{\partial V}{\partial y}, \phi \right\rangle + \frac{\dot{s}^{N,M}}{s^{N,M}} \left\langle y \frac{\partial}{\partial y} (V - P^N V), \phi \right\rangle + \frac{\dot{s}^{N,M}}{s^{N,M}} \left\langle y \frac{\partial \phi}{\partial y}, \phi \right\rangle \\
 & \quad + \langle \tilde{\rho}(V - \tilde{F}(\tilde{g})) - \rho^M(V - \tilde{F}(\tilde{g})), \phi \rangle \\
 & \quad + \langle \rho^M(V - \tilde{F}(\tilde{g})) - \rho^M(P^N V - \tilde{F}(\tilde{g})), \phi \rangle \\
 & \quad + \langle \rho^M(P^N V - \tilde{F}(\tilde{g})) - \rho^M(V^{N,M} - \tilde{F}(\tilde{g})), \phi \rangle \\
 & \quad + \langle \rho^M(V^{N,M} - \tilde{F}(\tilde{g})) - \rho^M(V^{N,M} - \tilde{F}(g^M)), \phi \rangle \\
 & \quad + \langle \hat{F}(s, \tilde{q}) - \hat{F}(s^{N,M}, q^M), \phi \rangle \\
 & \quad + \frac{1}{s} (\alpha_2^M - \tilde{\alpha}_2) V(t, 1) \phi(1) + \frac{1}{ss^{N,M}} (s - s^{N,M}) \alpha_2^M V(t, 1) \phi(1) \\
 & \quad - \frac{\alpha_2^M}{s^{N,M}} (V(t, 1) - P^N V(t, 1)) \phi(1) - \frac{\alpha_2^M}{s^{N,M}} (\phi(1))^2 \\
 & \quad + \frac{1}{s} (\tilde{h} - h^M) \phi(1) + \frac{1}{ss^{N,M}} (s^{N,M} - s) h^M \phi(1) + (s - s^{N,M})(\dot{s} - \dot{s}^{N,M}).
 \end{aligned}
 \tag{A.4}$$

Again, we use the fact that for any constant $c > 0$, we can write $\langle f, g \rangle \leq (c/2)\|f\|^2 + (1/2c)\|g\|^2$; moreover, the properties of $q \in \mathcal{Q}$ and (HN1) can be used to see that

$$\begin{aligned}
& \frac{d}{dt} (\|\phi\|^2 + (s - s^{N,M})^2) \\
& \leq \frac{1}{c\underline{s}^4} |\tilde{\mathcal{D}} - \mathcal{D}^M|_\infty^2 \left\| \frac{\partial V}{\partial y} \right\|^2 + c \left\| \frac{\partial \phi}{\partial y} \right\|^2 + \frac{4\bar{s}^2}{c\underline{s}^8} |s^{N,M} - s|^2 |\mathcal{D}^M|_\infty^2 \left\| \frac{\partial V}{\partial y} \right\|^2 \\
& \quad + c \left\| \frac{\partial \phi}{\partial y} \right\|^2 - \frac{2D_L}{\bar{s}^2} \left\| \frac{\partial \phi}{\partial y} \right\|^2 + \frac{1}{c\underline{s}^4} |\mathcal{D}^M|_\infty^2 \left\| \frac{\partial}{\partial y} (P^N V - V) \right\|^2 + c \left\| \frac{\partial \phi}{\partial y} \right\|^2 \\
& \quad + \frac{1}{c\underline{s}^2} |\tilde{\mathcal{V}}_1 - \mathcal{V}_1^M|_\infty^2 \|V\|^2 + c \left\| \frac{\partial \phi}{\partial y} \right\|^2 + \frac{1}{c\underline{s}^4} |s^{N,M} - s|^2 |\mathcal{V}_1^M|_\infty^2 \|V\|^2 \\
& \quad + c \left\| \frac{\partial \phi}{\partial y} \right\|^2 + \frac{1}{c\underline{s}^2} |\mathcal{V}_1^M|_\infty^2 \|\phi\|^2 + c \left\| \frac{\partial \phi}{\partial y} \right\|^2 + \frac{1}{c\underline{s}^2} |\mathcal{V}_1^M|_\infty^2 \|P^N V - V\|^2 \\
& \quad + c \left\| \frac{\partial \phi}{\partial y} \right\|^2 + \frac{1}{\underline{s}^2} \left\| \frac{\partial V}{\partial y} \right\|^2 |\tilde{\mathcal{V}}_2(s) - \mathcal{V}_2^M(s)|^2 + \|\phi\|^2 \\
& \quad + \frac{c(\mathcal{V}_2^{\text{Lip}})^2}{\underline{s}^2} \left\| \frac{\partial V}{\partial y} \right\|^2 |s - s^{N,M}|^2 + \frac{1}{c} \|\phi\|^2 + \frac{|\mathcal{V}_2^M(s^{N,M})|^2}{\underline{s}^4} |s^{N,M} - s|^2 \left\| \frac{\partial V}{\partial y} \right\|^2 \\
& \quad + \|\phi\|^2 + \frac{|\mathcal{V}_2^M(s^{N,M})|^2}{\underline{s}^2} \left\| \frac{\partial}{\partial y} (V - P^N V) \right\|^2 + \|\phi\|^2 + \frac{|\mathcal{V}_2^M(s^{N,M})|^2}{c\underline{s}^2} \|\phi\|^2 \\
& \quad + c \left\| \frac{\partial \phi}{\partial y} \right\|^2 + c \left(\frac{s}{s} - \frac{s^{N,M}}{s^{N,M}} \right)^2 \left\| \frac{\partial V}{\partial y} \right\|^2 + \frac{1}{c} \|\phi\|^2 + \frac{K_s^2}{\underline{s}^2} \left\| \frac{\partial}{\partial y} (V - P^N V) \right\|^2 \\
& \quad + \|\phi\|^2 + \frac{K_s^2}{c\underline{s}^2} \|\phi\|^2 + c \left\| \frac{\partial \phi}{\partial y} \right\|^2 + \|\tilde{\rho}(V - \tilde{F}(\tilde{g})) - \rho^M(V - \tilde{F}(\tilde{g}))\|^2 \\
& \quad + (\rho^{\text{Lip}})^2 \|V - P^N V\|^2 + (\rho^{\text{Lip}})^2 \|\phi\|^2 + (\rho^{\text{Lip}})^2 \|\tilde{F}(\tilde{g}) - \tilde{F}(g^M)\|^2 + 4\|\phi\|^2 \\
& \quad + 2\langle \hat{F}(s, \tilde{q}) - \hat{F}(s^{N,M}, q^M), \phi \rangle + \frac{|V(t, 1)|^2}{c\underline{s}^2} |\alpha_2^M - \tilde{\alpha}_2|^2 + c|\phi(1)|^2 \\
& \quad + \frac{(\alpha_2^M)^2 |V(t, 1)|^2}{c\underline{s}^4} |s - s^{N,M}|^2 + c|\phi(1)|^2 + \frac{(\alpha_2^M)^2}{c\underline{s}^2} |V(t, 1) - P^N V(t, 1)|^2 + c|\phi(1)|^2 \\
& \quad + \frac{1}{c\underline{s}^2} |\tilde{h} - h^M|^2 + c|\phi(1)|^2 + \frac{(h^M)^2}{c\underline{s}^4} |s^{N,M} - s|^2 + c|\phi(1)|^2 + \frac{1}{c} |s - s^{N,M}|^2 \\
& \quad + c|s - s^{N,M}|^2.
\end{aligned}$$

By assumption (HE1), there exist κ_1 and κ_2 such that $\kappa_1 = \sup_{t \in [0, T]} \|(\partial V / \partial y)\|^2$ and $\kappa_2 = \sup_{t \in [0, T]} \|V\|^2$. Under the hypothesis of parameter convergence, it follows that there exist constants \bar{D} , \bar{V}_1 , \bar{V}_2 , \bar{g} , \bar{h} , and $\bar{\alpha}$, independent of M such that $\sup_{t \in [0, T]} |\mathcal{D}^M|_\infty^2 \leq \bar{D}$, $\sup_{t \in [0, T]} |\mathcal{V}_1^M|_\infty^2 \leq \bar{V}_1$, $\sup_{t \in [0, T]} |\mathcal{V}_2^M(s^{N,M})|^2 \leq \bar{V}_2$, $\sup_{t \in [0, T]} (g^M)^2 \leq \bar{g}$, $\sup_{t \in [0, T]} (h^M)^2 \leq \bar{h}$, and $\sup_{t \in [0, T]} (\alpha_2^M)^2 \leq \bar{\alpha}$. Now, collecting the

terms containing \dot{s} and $\dot{s}^{N,M}$ (including those arising from the bound on $\langle \hat{F}(s, \tilde{q}) - \hat{F}(s^{N,M}, q^M), \phi \rangle$ derived in Lemma A.1), and using Lemma A.2, we obtain

$$\begin{aligned}
& \left(\frac{(\mathcal{V}_2^{\text{Lip}})^2}{\underline{s}^2} \left\| \frac{\partial V}{\partial y} \right\|^2 + 1 + 2C_7(g^M)^2 \right) c |\dot{s} - \dot{s}^{N,M}|^2 + \left(\left\| \frac{\partial V}{\partial y} \right\|^2 + 4(g^M)^2 \right) c \left(\frac{\dot{s}}{s} - \frac{\dot{s}^{N,M}}{s^{N,M}} \right)^2 \\
& \leq \left(\kappa_1 \left(\frac{\mathcal{V}_2^{\text{Lip}}}{\underline{s}} \right)^2 + 1 + 2C_7\bar{g} \right) c |\tilde{\mathcal{F}}(s, V - \tilde{F}(\tilde{g}), \Gamma) \\
& \quad - \tilde{\mathcal{F}}(s^{N,M}, p_{\Gamma^M}[V^{N,M} - \tilde{F}(g^M)], \Gamma^M)|^2 \\
& \quad + (\kappa_1 + 4\bar{g}) c \left| \frac{1}{s} \tilde{\mathcal{F}}(s, V - \tilde{F}(\tilde{g}), \Gamma) - \frac{1}{s^{N,M}} \tilde{\mathcal{F}}(s^{N,M}, p_{\Gamma^M}[V^{N,M} - \tilde{F}(g^M)], \Gamma^M) \right|^2 \\
& \leq \left(\kappa_1 \left(\frac{\mathcal{V}_2^{\text{Lip}}}{\underline{s}} \right)^2 + 1 + 2C_7\bar{g} \right) c \left\{ 2 |\tilde{\mathcal{F}}(s, V - \tilde{F}(\tilde{g}), \Gamma) - \tilde{\mathcal{F}}(s, V - \tilde{F}(\tilde{g}), \Gamma^M)|^2 \right. \\
& \quad + 4\mu^2 (|V - \tilde{F}(\tilde{g})|_\infty, |\Gamma^M|_\gamma) |s - s^{N,M}|^2 \\
& \quad \left. + 4\lambda^2 (s^{N,M}, |\Gamma^M|_\gamma) |V - \tilde{F}(\tilde{g}) - p_{\Gamma^M}[V^{N,M} - \tilde{F}(g^M)]|_\infty^2 \right\} \\
& \quad + \frac{2(\kappa_1 + 4\bar{g})}{\underline{s}^2} c \left\{ 2 |\tilde{\mathcal{F}}(s, V - \tilde{F}(\tilde{g}), \Gamma) - \tilde{\mathcal{F}}(s, V - \tilde{F}(\tilde{g}), \Gamma^M)|^2 \right. \\
& \quad + \left\{ 4\mu^2 (|V - \tilde{F}(\tilde{g})|_\infty, |\Gamma^M|_\gamma) + \left(\frac{K_s}{\underline{s}} \right)^2 \right\} |s - s^{N,M}|^2 \\
& \quad \left. + 4\lambda^2 (s^{N,M}, |\Gamma^M|_\gamma) |V - \tilde{F}(\tilde{g}) - p_{\Gamma^M}[V^{N,M} - \tilde{F}(g^M)]|_\infty^2 \right\}.
\end{aligned}$$

By assumption, the Γ^M 's range over a bounded (in the \mathcal{X}_γ topology) set, and by (HN1), $|s^{N,M}|$ is bounded uniformly in N and M ; therefore the continuity of μ and λ imply the existence of a set of constants C'_i such that

$$\begin{aligned}
& \left(\frac{(\mathcal{V}_2^{\text{Lip}})^2}{\underline{s}^2} \left\| \frac{\partial V}{\partial y} \right\|^2 + 1 + 2C_7(g^M)^2 \right) c |\dot{s} - \dot{s}^{N,M}|^2 + \left(\left\| \frac{\partial V}{\partial y} \right\|^2 + 4(g^M)^2 \right) c \left(\frac{\dot{s}}{s} - \frac{\dot{s}^{N,M}}{s^{N,M}} \right)^2 \\
& \leq C'_1 |\tilde{\mathcal{F}}(s, V - \tilde{F}(\tilde{g}), \Gamma) - \tilde{\mathcal{F}}(s, V - \tilde{F}(\tilde{g}), \Gamma^M)|^2 + C'_2 |s - s^{N,M}|^2 \\
& \quad + cC'_3 |V - \tilde{F}(\tilde{g}) - p_{\Gamma^M}[V^{N,M} - \tilde{F}(g^M)]|_\infty^2 \\
& \leq C'_1 |\tilde{\mathcal{F}}(s, V - \tilde{F}(\tilde{g}), \Gamma) - \tilde{\mathcal{F}}(s, V - \tilde{F}(\tilde{g}), \Gamma^M)|^2 + C'_2 |s - s^{N,M}|^2 \\
& \quad + 2cC'_3 |(V - \tilde{F}(\tilde{g})) - p_{\Gamma^M}[V - \tilde{F}(\tilde{g})]|_\infty^2 \\
& \quad + 8cC'_3 |V - P^N V|_\infty^2 + 8cC'_3 |\phi|_\infty^2 + 4cC'_3 |\tilde{F}(g^M) - \tilde{F}(\tilde{g})|_\infty^2 \\
& \leq C'_1 |\tilde{\mathcal{F}}(s, V - \tilde{F}(\tilde{g}), \Gamma) - \tilde{\mathcal{F}}(s, V - \tilde{F}(\tilde{g}), \Gamma^M)|^2 + C'_2 |s - s^{N,M}|^2 \\
& \quad + 2cC'_3 |(V - \tilde{F}(\tilde{g})) - p_{\Gamma^M}[V - \tilde{F}(\tilde{g})]|_\infty^2 \\
& \quad + 16cC'_3 \|V - P^N V\|^2 + 16cC'_3 \left\| \frac{\partial}{\partial y} (V - P^N V) \right\|^2 \\
& \quad + 16cC'_3 \|\phi\|^2 + 16cC'_3 \left\| \frac{\partial \phi}{\partial y} \right\|^2 + 4cC'_3 |\tilde{F}(g^M) - \tilde{F}(\tilde{g})|_\infty^2.
\end{aligned}$$

Incorporating the above estimate and the terms remaining from the estimate on $\langle \hat{F}(s, \tilde{q}) - \hat{F}(s^{N,M}, q^M), \phi \rangle$ with the previous estimate gives us a statement of the desired form, below:

$$(A.5) \quad \frac{d}{dt} (\|\phi\|^2 + (s - s^{N,M})^2) \leq \omega_1 \|\phi\|^2 + \omega_2 |s - s^{N,M}|^2 + \varepsilon \left\| \frac{\partial \phi}{\partial y} \right\|^2 + R_{\text{par}}^{N,M}(t) + R_{\text{app}}^{N,M}(t),$$

where we have defined the following quantities:

$$\begin{aligned} \omega_1 &\equiv \frac{\bar{V}_1 + \bar{V}_2 + K_s^2}{c\bar{s}^2} + 8 + \frac{2}{c} + (\rho^{\text{Lip}})^2 + 2C_5 + 10c + 16cC'_3, \\ \omega_2(t) &\equiv \frac{4\bar{s}^2 \bar{D}\kappa_1}{c\bar{s}^8} + \frac{\bar{V}_1 \kappa_2}{c\bar{s}^4} + \frac{\bar{V}_2 \kappa_1}{\bar{s}^4} + 2m_2(t) + 2\bar{g}m_3(t) + 2C_6 \bar{g} \bar{V}_2 \\ &\quad + \frac{2\bar{\alpha}(\kappa_1 + \kappa_2)}{c\bar{s}^4} + \frac{\bar{H}^2}{c\bar{s}^4} + \frac{1}{c} + C'_2, \\ \varepsilon &\equiv -\frac{2D_L}{\bar{s}^2} + 25c + 16cC'_3, \\ R_{\text{par}}^{N,M}(t) &\equiv \left(\frac{\kappa_1}{c\bar{s}^4} + 2\bar{g}C_2 \right) |\tilde{\mathcal{D}} - \mathcal{D}^M|_\infty^2 + \left(\frac{\kappa_2}{c\bar{s}^2} + 2\bar{g}C_3 \right) |\tilde{\mathcal{V}}_1 - \mathcal{V}_1^M|_\infty^2 \\ &\quad + \left(\frac{\kappa_1}{\bar{s}^2} + 2\bar{g}C_4 \right) |\tilde{\mathcal{V}}_2(s) - \mathcal{V}_2^M(s)|^2 + \|\tilde{\rho}(V - \tilde{F}(\tilde{g})) - \rho^M(V - \tilde{F}(\tilde{g}))\|^2 \\ &\quad + (\rho^{\text{Lip}})^2 \|\tilde{F}(\tilde{g}) - \tilde{F}(g^M)\|^2 + 2m_1(t) |g - g^M|^2 + 2C_1 |\dot{g} - \dot{g}^M|^2 \\ &\quad + \frac{2(\kappa_1 + \kappa_2)}{c\bar{s}^2} |\alpha_2^M - \tilde{\alpha}_2|^2 + \frac{1}{c\bar{s}^2} |\tilde{h} - h^M|^2 \\ &\quad + C'_1 |\tilde{\mathcal{F}}(s, V - \tilde{F}(\tilde{g}), \Gamma) - \tilde{\mathcal{F}}(s, V - \tilde{F}(\tilde{g}), \Gamma^M)|^2 \\ &\quad + 2cC'_3 |(V - \tilde{F}(\tilde{g})) - p_{\Gamma^M}[V - \tilde{F}(\tilde{g})]|_\infty^2 + 4cC'_3 \|\tilde{F}(\tilde{g}) - \tilde{F}(g^M)\|_\infty^2, \\ R_{\text{app}}^{N,M}(t) &\equiv \left\| \frac{\partial}{\partial y} (P^N V - V) \right\|^2 \left\{ \frac{\bar{D}}{c\bar{s}^4} + \frac{\bar{V}_2 + K_s^2}{\bar{s}^2} + \frac{2\bar{\alpha}}{c\bar{s}^2} + 16cC'_3 \right\} \\ &\quad + \|P^N V - V\|^2 \left\{ \frac{\bar{V}_1}{c\bar{s}^2} + (\rho^{\text{Lip}})^2 + \frac{2\bar{\alpha}}{c\bar{s}^2} + 16cC'_3 \right\}. \end{aligned}$$

Finally, we choose c so that $\varepsilon < 0$, obtaining the desired estimate

$$\frac{d}{dt} (\|P^N V - V^{N,M}\|^2 + |s - s^{N,M}|^2) \leq \omega(t) (\|P^N V - V^{N,M}\|^2 + |s - s^{N,M}|^2) + \beta^{N,M}(t),$$

where we define $\omega(t) = \omega_1 + \omega_2(t)$ and $\beta^{N,M}(t) = R_{\text{par}}^{N,M}(t) + R_{\text{app}}^{N,M}(t)$. Now the Gronwall inequality can be used to see that

$$\begin{aligned} (\|P^N V - V^{N,M}\|^2 + |s - s^{N,M}|^2) &\leq (\|P^N V - V^{N,M}\|^2 + |s - s^{N,M}|^2)|_{t=0} e^{\int_0^t \omega(s) ds} \\ &\quad + \int_0^t e^{\int_s^t \omega(\sigma) d\sigma} \beta^{N,M}(s) ds. \end{aligned}$$

At $t=0$, we have $\|P^N \tilde{V}_0 - P^N V_0^M\|^2 + |\tilde{s}_0 - s_0^M|^2 \leq \|\tilde{V}_0 - V_0^M\|^2 + |\tilde{s}_0 - s_0^M|^2 \rightarrow 0$ as $N, M \rightarrow \infty$, by the assumption of parameter convergence. It is easily verified that $\omega \in L^1(0, T)$; therefore it is clear that the first term in the above bound approaches zero as $N, M \rightarrow \infty$.

It is also easy to see that $\beta^{N,M}(t) \rightarrow 0$ as $N, M \rightarrow \infty$, for each t ; this follows from the assumption of parameter convergence, the continuity of \tilde{F} , and hypotheses (H \mathcal{F}), (P1), (P2), (HA1), and (HA2). Finally, it is straightforward to check the conditions for the Lebesgue dominated convergence theorem, and then we obtain the desired conclusion. \square

Proof of Theorem 5.4. We proceed as in the proof of Theorem 5.3, beginning with (A.4); however, we must treat differently the two boundary terms containing $\tilde{\alpha}_2$, α_2^M and \tilde{h} , h^M , now assumed to depend on \dot{s} and $\dot{s}^{N,M}$:

$$\begin{aligned}
 & \frac{1}{s} (\alpha_2^M - \tilde{\alpha}_2) V(t, 1) \phi(1) + \frac{1}{s} (\tilde{h} - h^M) \phi(1) \\
 &= \frac{1}{s} (\dot{s}^{N,M} - \dot{s}) V(t, 1) \phi(1) + \frac{1}{s} (-\tilde{\sigma} \dot{s} + \sigma^M \dot{s}^{N,M}) \phi(1) \\
 &= \frac{V(t, 1) \phi(1)}{s} (\dot{s}^{N,M} - \dot{s}) - \frac{\phi(1) \tilde{\sigma}}{s} (\dot{s} - \dot{s}^{N,M}) - \frac{\phi(1) \dot{s}^{N,M}}{s} (\tilde{\sigma} - \sigma^M) \\
 &\leq \frac{\phi(1)}{s} (V(t, 1) + \tilde{\sigma}) (\dot{s}^{N,M} - \dot{s}) + \frac{|\phi(1)| K_s}{s} |\tilde{\sigma} - \sigma^M| \\
 &\leq \frac{\phi(1)}{s} (V(t, 1) + \tilde{\sigma}) (\gamma^M(V^{N,M}(t, 1)) - \tilde{\gamma}(V(t, 1))) + \frac{c}{2} |\phi(1)|^2 \\
 &\quad + \frac{1}{2c} \left(\frac{K_s}{s} \right)^2 |\tilde{\sigma} - \sigma^M|^2 \\
 &= - \left(\frac{V(t, 1) + \tilde{\sigma}}{s} \right) \phi(1) (\gamma^M(P^N V(t, 1)) - \gamma^M(V^{N,M}(t, 1))) \\
 &\quad + \left(\frac{V(t, 1) + \tilde{\sigma}}{s} \right) \phi(1) \{ (\gamma^M(P^N V(t, 1)) - \gamma^M(V(t, 1))) \\
 &\quad + (\gamma^M(V(t, 1)) - \tilde{\gamma}(V(t, 1))) \} \\
 &\quad + \frac{c}{2} |\phi(1)|^2 + \frac{1}{2c} \left(\frac{K_s}{s} \right)^2 |\tilde{\sigma} - \sigma^M|^2 \\
 &\leq - \left(\frac{V(t, 1) + \tilde{\sigma}}{s} \right) \phi(1) (\gamma^M(P^N V(t, 1)) - \gamma^M(V^{N,M}(t, 1))) + \frac{c}{2} |\phi(1)|^2 \\
 &\quad + \frac{1}{c} \left(\frac{V(t, 1) + \tilde{\sigma}}{s} \right)^2 (|\gamma^M(P^N V(t, 1)) - \gamma^M(V(t, 1))|^2 + |\gamma^M(V(t, 1)) \\
 &\quad - \tilde{\gamma}(V(t, 1))|^2) \\
 &\quad + \frac{c}{2} |\phi(1)|^2 + \frac{1}{2c} \left(\frac{K_s}{s} \right)^2 |\tilde{\sigma} - \sigma^M|^2 \\
 &\leq - \left(\frac{V(t, 1) + \tilde{\sigma}}{s} \right) \phi(1) (\gamma^M(P^N V(t, 1)) - \gamma^M(V^{N,M}(t, 1))) \\
 &\quad + \frac{1}{c} \left(\frac{V(t, 1) + \tilde{\sigma}}{s} \right)^2 ((\gamma^{\text{Lip}})^2 |P^N V(t, 1) - V(t, 1)|^2 + |\gamma^M(V(t, 1)) - \tilde{\gamma}(V(t, 1))|^2) \\
 &\quad + c |\phi(1)|^2 + \frac{1}{2c} \left(\frac{K_s}{s} \right)^2 |\tilde{\sigma} - \sigma^M|^2.
 \end{aligned}$$

Since $V(t, 1) = U(t, 1) \geq 0$, $\tilde{\sigma} \in \mathfrak{H}$ implies that $\tilde{\sigma} \geq 0$, and $\tilde{\gamma}^M \in \mathcal{G}$ implies that $\tilde{\gamma}^M$ is a nondecreasing function for every M , we can conclude that the first term in the above inequality is negative and can therefore be dropped. We can then conclude that for some constants \bar{C}_i , which are independent of N and M , it follows that

$$\begin{aligned} & \frac{1}{s} (\alpha_2^M - \tilde{\alpha}_2) V(t, 1) \phi(1) + \frac{1}{s} (\tilde{h} - h^M) \phi(1) \\ & \leq \bar{C}_1 \|P^N V - V\|^2 + \bar{C}_2 \left\| \frac{\partial}{\partial y} (P^N V - V) \right\|^2 + \bar{C}_3 |\tilde{\gamma}^M(V(t, 1)) - \tilde{\gamma}(V(t, 1))|^2 \\ & \quad + c |\phi(1)|^2 + \bar{C}_4 |\tilde{\sigma} - \sigma^M|^2. \end{aligned}$$

These terms would then be incorporated into expression (A.5), with the appropriate changes in the definitions of ω_1 , ω_2 , ε , $R_{\text{par}}^{N,M}$, and $R_{\text{app}}^{N,M}$, and then ω and $\beta^{N,M}$. The proof follows exactly as that of Theorem 5.3. \square

REFERENCES

- [1] J. ALBRECHT, L. COLLATZ, AND K.-H. HOFFMANN, EDS., *Numerical Treatment of Free Boundary Value Problems*, Birkhäuser-Verlag, Boston, 1982.
- [2] H. T. BANKS, *On a variational approach to some parameter estimation problems*, in *Lecture Notes Control Inform. Sci.*, Vol. 75, Springer-Verlag, New York, Berlin, 1985, pp. 1-23.
- [3] H. T. BANKS, J. M. CROWLEY, AND K. KUNISCH, *Cubic spline approximation techniques for parameter estimation in distributed systems*, *IEEE Trans. Automat. Control*, 28(1983), pp. 773-786.
- [4] H. T. BANKS AND P. D. LAMM, *Estimation of variable coefficients in parabolic distributed systems*, *IEEE Trans. Automat. Control*, 30(1985), pp. 386-398.
- [5] H. T. BANKS AND K. KUNISCH, *Estimation Techniques for Distributed Systems*, Birkhäuser, Boston, 1989.
- [6] H. T. BANKS AND K. A. MURPHY, *Estimation of nonlinearities in parabolic models for growth, predation and dispersal of populations*, *J. Math. Anal. Appl.*, 141(1989), pp. 580-602.
- [7] A. BOSSAVIT, A. DAMLAMIAN, AND M. FREMOND, EDS., *Free Boundary Problems: Applications and Theory*, Vol. III, *Research Notes in Mathematics*, 120, Pitman, Boston, 1985.
- [8] H. T. CHANG AND B. E. RITTMAN, *Mathematical modeling of biofilm on activated carbon*, *Environ. Sci. Tech.*, 21(1987), pp. 273-280.
- [9] 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., Akademik-Verlag, Berlin, 1979, pp. 45-58.
- [10] J. CRANK, *Free and Moving Boundary Problems*, Clarendon Press, Oxford, UK, 1984.
- [11] J. F. EPPERSON, *Finite element methods for a class of nonlinear evolution equations*, *SIAM J. Numer. Anal.*, 21(1984), pp. 1066-1079.
- [12] A. FASANO AND M. PRIMICERIO, EDS., *Free Boundary Problems: Theory and Applications*, Vol. II, *Research Notes in Mathematics*, 79, Pitman, Boston, 1983.
- [13] A. FASANO AND R. RICCI, *Penetration of solvents into glassy polymers*, in *Free Boundary Problems: Applications and Theory*, Vol. III, *Research Notes in Mathematics*, 120, Pitman, Boston, 1985, pp. 132-139.
- [14] A. FRIEDMAN, *Variational Principles and Free-Boundary Problems*, Wiley-Interscience, New York, 1982.
- [15] K.-H. HOFFMANN AND H.-J. KORNSTAEDT, *Ein numerisches verfahren zur losung eines identifizierungs-problems bei der wärmeleitungsgleichung*, in *Numerical Treatment of Free Boundary Value Problems*, Birkhäuser-Verlag, Boston, 1982, pp. 108-126.
- [16] K.-H. HOFFMANN AND M. NIEZGODGKA, *Control of parabolic systems involving free boundaries*, in *Free Boundary Problems: Theory and Applications*, Vol. II, *Research Notes in Mathematics*, 79, Pitman, Boston, 1983, pp. 431-462.
- [17] P. JOCHUM, *The numerical solution of the inverse Stefan problem*, *Numer. Math.*, 34(1980), pp. 411-429.
- [18] C. KRAVARIS AND J. H. SEINFELD, *Identification of parameters in distributed parameter systems by regularization*, *SIAM J. Control Optim.*, 23(1985), pp. 217-241.
- [19] P. K. LAMM AND K. A. MURPHY, *Estimation of discontinuous coefficients and boundary parameters for hyperbolic systems*, *Quart. J. Appl. Math.*, 46(1988), pp. 1-22.

- [20] C. H. LI, *A finite-element front-tracking enthalpy method for Stefan problems*, IMA J. Numer. Anal., 3(1983), pp. 87–107.
- [21] A. I. LIAPIS, G. G. LIPSCOMB, AND O. K. CROSSER, *A model of oxygen diffusion in absorbing tissue*, Math. Model., 3(1982), pp. 83–92.
- [22] G. H. MEYER, *One-dimensional parabolic free boundary problems*, SIAM Rev., 19(1977), pp. 17–34.
- [23] K. A. MURPHY, *Estimation of time- and state-dependent delays and other parameters in functional differential equations*, SIAM J. Appl. Math., 50 (1990), pp. 972–1000.
- [24] ———, *Parameter estimation in moving boundary problems*, Appl. Math. Lett., 1(1988), pp. 303–306.
- [25] ———, *Parameter estimation in moving boundary problems*, in Proc. 27th IEEE Conf. on Decision and Control, Austin, TX, December 7–9, 1988, pp. 1656–1661.
- [26] A. S. PERELSON AND E. A. COUTSIAS, *A moving boundary model of acrosomal elongation*, J. Math. Biol., 23(1986), pp. 361–379.
- [27] P. M. PRENTER, *Splines and Variational Methods*, Wiley-Interscience, New York, 1975.
- [28] M. H. SCHULTZ, *Spline Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1973.
- [29] D. G. WILSON, A. D. SOLOMON, AND P. T. BOGGS, EDS., *Moving Boundary Problems*, Academic Press, New York, 1978.

A LINEAR PROGRAMMING APPROACH TO THE SEARCH GAME ON A NETWORK WITH MOBILE HIDER*

EDWARD J. ANDERSON[†] AND MIGUEL ARAMENDIA[†]

Abstract. This paper discusses a search game on a network Q with two players, a searcher and a hider, who each move with continuous trajectories starting from different points subject to a maximal speed and termination time T . This is a zero-sum game with payoff given by the time elapsed until the searcher reaches a point that is occupied by the hider (if this happens), and T otherwise. The problem is formulated as an infinite-dimensional linear program, and the extreme points and the reduced cost functional are studied. An algorithm is derived for this problem, and how it works on an example is demonstrated.

Key words. search game, infinite-dimensional program, extreme point

AMS(MOS) subject classifications. 90D26, 49A45, 90B40, 90C48

1. Introduction. We are concerned in this paper with the following search game, which was originally suggested by Isaacs [9]. The search space Q is a connected network of finite length, with no crossovers except at the nodes, and two distinguished nodes, denoted by O_s and O_h , known by both players, the searcher and the hider. A pure hider strategy and a pure searcher strategy are continuous trajectories that, starting at O_s and O_h , respectively, at the time 0, move along Q with speed less than or equal to 1 until time T , where T is the termination time of the game. Neither player can see the other. The capture time, which the searcher seeks to minimize and the hider to maximize, is the time elapsed until the searcher reaches the point that is occupied by the hider (if this occurs), and T otherwise. This scheme defines a two-person zero-sum game with the capture time as the payoff function.

As a stepping stone to this problem, Isaacs proposed the simpler problem where Q is the boundary of a circle. This game was solved by Alpern [1], Foreman [6], and Zelikin [11] with different formulations of the initial conditions. Some versions of this game with a fixed termination time were considered by Foreman [7]. These authors give results in terms of the mixed strategy, called the *cohato* strategy (short for coin-half-tour) that consists of oscillating at speed 1 between a given point and its antipode, each time choosing equiprobably between the counterclockwise and the clockwise directions.

The search game for the network consisting of two nodes connected by three arcs of equal length has been solved by Alpern and Asic [3]. The value of the game is 3, and the strategy considered for the searcher randomly oscillates at speed 1 between the initial point and its antipode. The strategy for the hider is similar, except when the same arc occurs consecutively; he waits a small distance δ from the intervening node for a period of length 2δ and then resumes the oscillation strategy. For both games, the hider does not have an optimal solution, but can achieve an ε -optimal solution.

Gal [8] proved that the search game on a network has a value, say V , and conjectured that $\Gamma \leq V \leq 2\Gamma$, where Γ is the length of the network. A counterexample to the conjectured upper bound was given by Fitzgerald [5]. He proved that for any positive number k , there exists a network with length Γ with value greater than $k\Gamma$.

* Received by the editors May 23, 1990; accepted for publication (in revised form) March 7, 1991.

[†] Management Studies Group, Department of Engineering, University of Cambridge, Cambridge CB2 1RX, England.

The conjectured lower bound is also incorrect, as has been shown by Alpern and Asic [2] for the figure-eight network.

Our contribution is the study of the linear programming problem to which this search game leads. We construct an algorithm for the solution of this search game and give one example of its performance.

2. Formulation of the game. Formally, we consider Q with the topology induced by the distance d , where $d(x, y)$ is defined as the length of the shortest path in Q between x and y . We say that a trajectory s in Q has maximal speed 1 if

$$d(s(t_1), s(t_2)) \leq |t_1 - t_2|, \quad \text{for all } t_1, t_2 \in [0, T].$$

It is clear that if s has maximal speed 1, it is continuous on $[0, T]$. The set of pure searcher strategies and hider strategies are defined as follows:

$$TS = \{s \in C([0, T], Q) : s(0) = O_s, d(s(t_1), s(t_2)) \leq |t_1 - t_2|, \text{ for all } t_1, t_2 \in [0, T]\},$$

$$TH = \{h \in C([0, T], Q) : h(0) = O_h, d(h(t_1), h(t_2)) \leq |t_1 - t_2|, \text{ for all } t_1, t_2 \in [0, T]\},$$

where $C([0, T], Q)$ is the set of all continuous functions defined on $[0, T]$ with values in Q . We consider TS and TH endowed with the inherited topology from $C([0, T], Q)$ defined by the distance

$$d(f, g) = \max \{d(f(t), g(t)) : t \in [0, T]\}.$$

TS and TH are obviously Hausdorff spaces, and it is easy to prove that they are compact using the Ascoli theorem.

The payoff function for a pure searcher strategy s and a pure hider strategy h is defined as follows:

$$K(s, h) = \begin{cases} \min \{t : s(t) = h(t), t \in [0, T]\} & \text{if it exists,} \\ T & \text{otherwise.} \end{cases}$$

It is worth noting that if there is a $t \in [0, T]$ such that $s(t) = h(t)$, then since the function $f(t) = d(s(t), h(t))$ is continuous on $[0, T]$, the set $\{t \in [0, T] : s(t) = h(t)\} = f^{-1}(\{0\})$ is compact, and therefore $K(s, h)$ is well defined. Moreover, if $\hat{t} = K(s, h) < T$, then $s(\hat{t}) = h(\hat{t})$. It is easy to see (Gal [8]) that for any fixed $h \in TH$, $K(s, h)$ is lower semicontinuous on TS , and for any fixed $s \in TS$, $K(s, h)$ is lower semicontinuous on TH .

A mixed strategy $\mu(\nu)$ for the searcher (hider) is a regular Borel probability measure on $TS(TH)$. We denote by $P(TS)(P(TH))$ the set of all mixed strategies. The value of the payoff function for the mixed strategies $\mu \in P(TS)$ and $\nu \in P(TH)$ is

$$\begin{aligned} \int_{TS \times TH} K(s, h) d(\mu(s), \nu(h)) &= \int_{TS} \int_{TH} K(s, h) d\nu(h) d\mu(s) \\ &= \int_{TH} \int_{TS} K(s, h) d\mu(s) d\nu(h). \end{aligned}$$

We write $K(s, \nu)$ for $\int_{TH} K(s, h) d\nu(h)$ and $K(\mu, h)$ for $\int_{TS} K(s, h) d\mu(s)$.

3. The associated linear programs. We next consider the two linear programming problems that this dynamic search game yields. In the game, if the hider chooses the strategy ν , then he is certain of obtaining at least $V(\nu) = \inf \{K(\mu, \nu) : \mu \in P(TS)\}$. Thus, since

$$\begin{aligned} K(\mu, \nu) &= \int_{TS} K(s, \nu) d\mu(s) \\ &\geq \inf \{K(s, \nu) : s \in TS\}, \end{aligned}$$

we have

$$V(\nu) = \inf \{K(s, \nu) : s \in TS\}.$$

The hider wishes to choose ν to make $V(\nu)$ as large as possible. This problem can be expressed as the following linear programming problem:

$$\begin{aligned} & \text{maximize} && \alpha \\ & \text{subject to} && \int_{TH} K(s, h) d\nu(h) \geq \alpha, \quad \text{for all } s \in TS, \\ & && \nu \in P(TH), \quad \alpha \in R. \end{aligned}$$

Clearly, the optimal value of the problem is the value of the game. Since the value is positive, taking $\nu' = (1/\alpha)\nu$, and since $\int_{TH} d\nu'(h) = (1/\alpha) \int_{TH} d\nu(h) = 1/\alpha$, the problem becomes

$$\begin{aligned} \text{(HLP1)} \quad & \text{minimize} && \int_{TH} d\nu'(h) \\ & \text{subject to} && \int_{TH} K(s, h) d\nu'(h) \geq 1, \quad \text{for all } s \in TS, \\ & && \nu' \geq 0, \quad \nu' \in M(TH), \end{aligned}$$

where $M(TH)$ denotes the set of all regular Borel measures on TH .

Similarly, for the searcher we obtain the problem

$$\begin{aligned} \text{(SLP1)} \quad & \text{maximize} && \int_{TS} d\mu(s) \\ & \text{subject to} && \int_{TS} K(s, h) d\mu(s) \leq 1, \quad \text{for all } h \in TH, \\ & && \mu \geq 0, \quad \mu \in M(TS), \end{aligned}$$

where $M(TS)$ denotes the set of all regular Borel measures on TS .

Following the framework of the duality theory described in Anderson and Nash [4], we see that these problems are dual to each other with respect to the dual pairs $(M(TS), F(TS))$ and $(M(TH), F(TH))$, where $F(TS)$ and $F(TH)$ are the spaces of all real-valued functions defined on TS and TH , respectively, which are integrable for all regular Borel measures on TS and TH , respectively. Thus the weak duality theorem implies that if μ is a feasible point of (SLP1) and ν is a feasible point of (HLP1), then $\int_{TS} d\mu(s) \leq \int_{TH} d\nu(h)$. A special feature of these two problems is that there is no duality gap since both have the same value (precisely $1/V$ with V the value of the game).

4. Feasible points. The duality of (SLP1) and (HLP1) means that we could approach the solution of the problem by seeking an algorithm for either version. However, a necessary part of such an algorithm is an effective procedure for checking the feasibility of a solution. This proves much harder for (SLP1) than for (HLP1). Consequently, we will now concentrate on the hider linear program.

Introducing a slack variable z , the problem (HLP1) becomes

$$\begin{aligned} \text{(HLP2)} \quad & \text{minimize} && \int_{TH} d\nu(h) \\ & \text{subject to} && \int_{TH} K(s, h) d\nu(h) - z(s) = 1, \quad \text{for all } s \in TS, \\ & && \nu \geq 0, \quad \nu \in M(TH), \quad z \geq 0, \quad z \in F(TS), \end{aligned}$$

where, as before, $M(TH)$ denotes the set of all regular Borel measures on TH , and $F(TS)$ denotes the set of all real-valued functions defined on TS , which are integrable for all regular Borel measures on TS . By definition, the feasible region of (HLP2) is the set

$$\{(\nu, z) \in M(TH) \times F(TS): z(s) = \int_{TH} K(s, h) d\nu(h) - 1 \geq 0, \nu \geq 0\}.$$

We say that ν is a feasible point if (ν, z) is a feasible point of the feasible region of (HLP2).

Our attention is focused on atomic measures. These correspond to mixed strategies in which the hider selects from a finite number of possible trajectories with some given probability distribution. Let ν be an atomic measure concentrated on h_1, \dots, h_n with masses $\lambda_1, \dots, \lambda_n$, then ν is a feasible point if $\lambda_i \geq 0$, for $i = 1, \dots, n$, and

$$z(s) = \lambda_1 K(s, h_1) + \dots + \lambda_n K(s, h_n) - 1 \geq 0, \text{ for all } s \in TS.$$

Since the masses are nonnegative and $K(s, h_i)$ is lower semicontinuous on TS , the slack variable z is lower semicontinuous on the compact space TS and therefore has a minimum on TS . We will show that this minimum occurs in a given finite set independently of the masses λ_i .

To begin, we suppose that ν is an atomic measure concentrated on a single point h_1 with mass λ_1 . We wish to define the searcher trajectory that meets h_1 quickest. Let g be defined on $[0, T]$ as $g(t) = t - d(O_s, h_1(t))$. Then $g(0) \leq 0$ and $g(T) > 0$ (we assume that T is greater than the maximal distance between points of Q). Thus, since g is continuous, there is at least one t where the function g takes the value zero. Let t_1 be the first time at which $h_1(t)$ is a distance t from O_s , i.e., $t_1 = \min \{t: t \in [0, T], g(t) = 0\}$.

We define the trajectory s_1 that, starting at the point O_s , moves along the shortest path between O_s and $h_1(t_1)$, or one such path, until it reaches the point $h_1(t_1)$ and then moves arbitrarily along the network always with speed 1. Note that $s_1(t_1) = h_1(t_1)$. We denote by $S(O_s; h_1)$ the set with the single element s_1 .

LEMMA 4.1. *Let ν be an atomic measure concentrated on h_1 with mass $\lambda_1 \geq 0$. Then the associated slack variable z satisfies $z(s_1) \leq z(s)$, for all $s \in TS$, where s_1 is the pure search strategy of $S(O_s; h_1)$.*

Proof. Take s as any element of TS . Let $\hat{t}_1 = K(s, h_1)$. Since s has speed less than or equal to 1, we have $d(h_1(\hat{t}_1), O_s) = d(s(\hat{t}_1), s(0)) \leq \hat{t}_1$. Therefore $g(\hat{t}_1) \geq 0$ and, from the definition of t_1 , $\hat{t}_1 \geq t_1$. Therefore $K(s, h_1) \geq K(s_1, h_1)$, and the theorem is proved. \square

To generalize this theorem to atomic measures concentrated on more than one point, we consider the set $S(O_s; h_1, \dots, h_n)$ defined as follows:

$$S(O_s; h_1, \dots, h_n) = \{s_\pi: \pi \in \Pi\},$$

where each s_π is a pure search strategy defined by repeating the above process with the n trajectories h_i . In fact, the trajectory s_π hunts the n trajectories h_i following the order given by the permutation π , and we call this set of search strategies the hunting set for h_1, \dots, h_n . For instance, the trajectory $s_{1, \dots, n}$ is formally defined as follows: let t_i be the value defined by

$$t_i = \min \{T, \{t: t \in [t_{i-1}, T], t - t_{i-1} = d(h_{i-1}(t_{i-1}), h_i(t))\}\}, \text{ for } i = 1, \dots, n,$$

where $t_0 = 0$ and $h_0(t_0) = O_s$. Let $f_i: [0, d(h_{i-1}(t_{i-1}), h_i(t_i))] \rightarrow Q$ be the canonical map

of a shortest path between $h_{i-1}(t_{i-1})$ and $h_i(t_i)$, for $i = 1, \dots, n$. Then

$$\begin{aligned} s_{1,\dots,n}(t) &= f_1(t), & \text{for all } t: 0 \leq t \leq t_1, \\ s_{1,\dots,n}(t) &= f_2(t - t_1), & \text{for all } t: t_1 < t \leq t_2, \\ &\vdots & \vdots \\ s_{1,\dots,n}(t) &= f_n(t - t_{n-1}), & \text{for all } t: t_{n-1} < t \leq t_n \end{aligned}$$

if $t_n < T$, $s_{1,\dots,n}$ keeps moving from $h_n(t_n)$ to the end of the arc in the same direction as before and then moves arbitrarily from node to node with speed 1 until time T runs out. Note that $s_{1,\dots,n}(0) = O_s$, its speed is always 1, and if $t_i < T$ then $s_{1,\dots,n}(t_i) = h_i(t_i)$.

THEOREM 4.2. *Let ν be an atomic measure concentrated on h_1, \dots, h_n with nonnegative masses $\lambda_1, \dots, \lambda_n$. Then the associated slack variable z achieves its minimum at an element of $S(O_s; h_1, \dots, h_n)$.*

Proof. Take s to be any element of TS . We suppose without loss of generality that $K(s, h_1) \leq K(s, h_2) \leq \dots \leq K(s, h_n)$, and we prove that $K(s_{1,\dots,n}, h_i) \leq K(s, h_i)$ for $i = 1, \dots, n$. Since $\lambda_i \geq 0$, the result follows.

From Theorem 4.1, we have $K(s_{1,\dots,n}, h_1) = t_1 \leq K(s, h_1)$. Using induction on k , we assume that the trajectory $s_{1,\dots,n}$ satisfies

$$(1) \quad K(s_{1,\dots,n}, h_i) \leq K(s, h_i), \quad \text{for } i = 1, \dots, k-1,$$

and we will prove that (1) holds for $i = k$.

It is clear that $K(s_{1,\dots,n}, h_k) \leq t_k$. A strict inequality may arise when $s_{1,\dots,n}$ hits h_k prior to hitting h_{k-1} . Let $\hat{t}_k = K(s, h_k)$ and $\hat{t}_{k-1} = K(s, h_{k-1})$. Since the speed of h_{k-1} is less than or equal to 1, we have

$$d(h_{k-1}(t), h_{k-1}(t_{k-1})) \leq t - t_{k-1}, \quad \text{for } t \geq t_{k-1},$$

in particular for $t = \hat{t}_{k-1}$, and since $s(\hat{t}_{k-1}) = h_{k-1}(\hat{t}_{k-1})$ (otherwise $\hat{t}_{k-1} = T$ and then (1) clearly holds for $i = k$), it follows that $d(s(\hat{t}_{k-1}), h_{k-1}(t_{k-1})) \leq \hat{t}_{k-1} - t_{k-1}$. Thus

$$\begin{aligned} d(s(t), h_{k-1}(t_{k-1})) &\leq d(s(t), s(\hat{t}_{k-1})) + d(s(\hat{t}_{k-1}), h_{k-1}(t_{k-1})) \\ &\leq t - \hat{t}_{k-1} + \hat{t}_{k-1} - t_{k-1} = t - t_{k-1}, \quad \text{for } t \geq \hat{t}_{k-1}, \end{aligned}$$

in particular for $t = \hat{t}_k$, and since $s(\hat{t}_k) = h_k(\hat{t}_k)$, it follows that $d(h_k(\hat{t}_k), h_{k-1}(t_{k-1})) \leq \hat{t}_k - t_{k-1}$. Hence the real-valued continuous function $g(t) = t - t_{k-1} - d(h_{k-1}(t_{k-1}), h_k(t))$ satisfies $g(t_{k-1}) \leq 0$ and $g(\hat{t}_k) \geq 0$. Thus, from the definition of t_k , $\hat{t}_k \geq t_k$, and the theorem is proved. \square

COROLLARY 4.3. *An atomic measure ν concentrated on h_1, \dots, h_n with masses $\lambda_1, \dots, \lambda_n$ and slack variable z is a feasible solution of (HLP2) if and only if*

$$\begin{aligned} \lambda_i &\geq 0, \quad \text{for } i = 1, \dots, n, \quad \text{and} \\ z(s) &\geq 0, \quad \text{for all } s \in S(O_s; h_1, \dots, h_n). \end{aligned}$$

This result gives an easily computable check for feasibility. It turns out that the feasibility check of Corollary 4.3 is a vital tool in the development of an algorithm for the hidden linear program.

5. Extreme points. Let ν be a feasible point of (HLP2) concentrated on h_1, \dots, h_n with masses $\lambda_1, \dots, \lambda_n$ and slack variable z . We say that ν is an extreme atomic measure if (ν, z) is an extreme point of the feasible region of (HLP2). Following Anderson and Nash [4], ν is an extreme point if and only if $B(\nu) \times B(z) \cap N(A) = \{0\}$,

where

$$\begin{aligned} B(z) &= \{y \in F(TS): z \pm \varepsilon y \geq 0 \text{ for some } \varepsilon \in R, \varepsilon > 0\}, \\ B(\nu) &= \{\mu \in M(TH): \nu \pm \varepsilon \mu \geq 0 \text{ for some } \varepsilon \in R, \varepsilon > 0\} \\ &\subset \{\mu \in M(TH): \mu \text{ concentrated on } h_1, \dots, h_n\}, \\ N(A) &= \left\{ (\mu, y) \in M(TH) \times F(TS): \int_{TH} K(s, h) d\mu(h) = y(s) \text{ for all } s \in TS \right\}. \end{aligned}$$

Letting $\text{constr}(\nu) = \{s \in TS: z(s) = 0\}$, we have the following result.

THEOREM 5.1. *Let ν be a feasible point of (HLP2) concentrated on h_1, \dots, h_n with masses $\lambda_1, \dots, \lambda_n$. If ν is an extreme point, then $\text{constr}(\nu)$ is nonempty.*

Proof. Suppose otherwise and let η be the minimum value of z on TS . By hypothesis, $\eta > 0$. Let k be an index such that $\lambda_k > 0$ ($\nu = 0$ is not feasible). The function $y(s) = K(s, h_k)$ satisfies $y \in B(z)$ (with $\varepsilon = \eta/T$). If we define μ to be the measure concentrated on h_k with mass 1, then $(\mu, \nu) \in B(\nu) \times B(z) \cap N(A)$. \square

From now on, we suppose that $\text{constr}(\nu) \neq \emptyset$. The fact that the minimum of the slack variable z is achieved at an element of $S(O_s; h_1, \dots, h_n)$ implies that the intersection of both sets is nonempty. Let $S(O_s; h_1, \dots, h_n) \cap \text{constr}(\nu) = \{\hat{s}_1, \dots, \hat{s}_m\}$. We define the matrix A_ν by

$$A_\nu = \begin{pmatrix} K(\hat{s}_1, h_1) & \cdots & K(\hat{s}_1, h_n) \\ \vdots & \ddots & \vdots \\ K(\hat{s}_m, h_1) & \cdots & K(\hat{s}_m, h_n) \end{pmatrix}.$$

This matrix is made up of meeting times between the pure hider strategies h_i and certain search trajectories in the hunting set for h_1, \dots, h_n . In these circumstances, we have the following straightforward characterization of ν as an extreme point when the masses are greater than zero.

THEOREM 5.2. *Let ν be a feasible point concentrated on h_1, \dots, h_n with positive masses $\lambda_1, \dots, \lambda_n$. Then ν is an extreme point if and only if $\text{rank}(A_\nu) = n$.*

Proof. Suppose first that $\text{rank}(A_\nu) < n$. Then there is a nonzero $\alpha \in R^n$ with $A_\nu \alpha = 0$. We consider the atomic measures $\mu = \sum_{i=1}^n \alpha_i \delta_{h_i}$, $\nu_{+\varepsilon} = \nu + \varepsilon \mu$, and $\nu_{-\varepsilon} = \nu - \varepsilon \mu$, where δ_h denotes the measure with mass 1 concentrated at point h . We will show that if ε is chosen small enough, both of these points are feasible. We define the value ε_1 as follows:

$$\varepsilon_1 = \min \left\{ \frac{\lambda_i}{|\alpha_i|}: \alpha_i \neq 0, i = 1, \dots, n \right\}.$$

Thus $0 < \varepsilon_1 < +\infty$, and $\nu \pm \varepsilon \mu \geq 0$, for all $\varepsilon: 0 \leq \varepsilon \leq \varepsilon_1$. We define the value ε_2 as follows:

$$\varepsilon_2 = \min \left\{ \frac{z(s)}{|\sum_{i=1}^n \alpha_i K(s, h_i)|} \right\},$$

where the minimum is taken over $s \in S(O_s; h_1, \dots, h_n)$ such that the denominator is not zero, and if there is no such s , then we take $\varepsilon_2 = +\infty$. Note that ε_2 is positive since if there is an $\hat{s} \in S(O_s; h_1, \dots, h_n)$ such that the numerator is zero, then $\hat{s} \in \text{constr}(\nu)$ and therefore $\sum_{i=1}^n \alpha_i K(\hat{s}, h_i) = 0$. Let $\nu_{+\varepsilon}$ and $\nu_{-\varepsilon}$ be the associated slack variables of

$\nu_{+\varepsilon}$ and $\nu_{-\varepsilon}$, respectively. We have

$$\begin{aligned} z_{\pm\varepsilon}(s) &= (\lambda_1 \pm \varepsilon \alpha_1)K(s, h_1) + \cdots + (\lambda_n \pm \varepsilon \alpha_n)K(s, h_n) - 1 \\ &= z(s) \pm \varepsilon \sum_{i=1}^n \alpha_i K(s, h_i) \\ &\geq 0, \quad \text{for all } s \in S(O_s; h_1, \dots, h_n), \text{ for all } \varepsilon: 0 \leq \varepsilon \leq \varepsilon_2. \end{aligned}$$

So, since the minimum of the slack variable $z_{\pm\varepsilon}$ on TS is reached at an element of $S(O_s; h_1, \dots, h_n)$ when the masses are nonnegative, we have $z_{\pm\varepsilon}(s) \geq 0$, for all $s \in TS$, for all $\varepsilon: 0 \leq \varepsilon \leq \min\{\varepsilon_1, \varepsilon_2\}$. Therefore ν is not an extreme point.

Now suppose that $\text{rank}(A_\nu) = n$. Let $(\mu, y) \in B(\nu) \times B(z) \cap N(A)$; then μ is an atomic measure concentrated on h_1, \dots, h_n with masses $\alpha_1, \dots, \alpha_n$ and $y(s) = \sum_{i=1}^n \alpha_i K(s, h_i)$. From the definition of $B(z)$, it follows that $y(s) = 0$ for all $s \in \text{constr}(\nu)$; therefore, as $\{\hat{s}_1, \dots, \hat{s}_m\} = S(O_s; h_1, \dots, h_n) \cap \text{constr}(\nu)$, it follows that

$$\sum_{i=1}^n \alpha_i K(\hat{s}_j, h_i) = 0, \quad \text{for } j = 1, \dots, m.$$

However, since $\text{rank}(A_\nu) = n$, α is identically zero and hence $B(\nu) \times B(z) \cap N(A) = \{0\}$. \square

It is clear from this theorem that no extreme atomic measure has mass at point O_s . Next, we extend a part of the previous theorem to the case where the masses are nonnegative and the searcher trajectories are not necessarily elements of the hunting set. First, we define the payoff matrix generated by $\{h_1, \dots, h_n\}$ and $\{s_1, \dots, s_n\}$ to be the matrix

$$\begin{pmatrix} K(s_1, h_1) & \cdots & K(s_1, h_n) \\ \vdots & \ddots & \vdots \\ K(s_n, h_1) & \cdots & K(s_n, h_n) \end{pmatrix}.$$

THEOREM 5.3. *Let ν be a feasible point concentrated on h_1, \dots, h_n with masses $\lambda_1, \dots, \lambda_n$. Let s_1, \dots, s_n be n searcher trajectories of the set $\text{constr}(\nu)$. If the payoff matrix A generated by $\{h_1, \dots, h_n\}$ and $\{s_1, \dots, s_n\}$ has rank n , then ν is an extreme point. Moreover, $\lambda^T = (\lambda_1, \dots, \lambda_n)$ satisfies $\lambda = A^{-1}e$, where $e^T = (1, \dots, 1)$.*

Proof. The first part is proved in the same way as the above proof. Since the searcher trajectories belong to the set $\text{constr}(\nu)$, we have

$$\sum_{i=1}^n \lambda_i K(s_j, h_i) = 1, \quad \text{for all } j = 1, \dots, n.$$

Thus, using the invertibility of A , the result follows. \square

It is worth noting that this result includes the case of atomic measures with zero masses at some of the points h_i . Later, it will be convenient to consider such solutions.

6. Purification algorithm. Theorem 5.2 raises the question of what to do when the rank of A is less than n . In this section we consider how to construct an extreme atomic measure in this case, which may arise during the course of an algorithm based on extreme points. In fact, we describe a purification algorithm that, starting from an initial feasible atomic measure, obtains an improved extreme atomic measure. This algorithm is based on the purification algorithm developed by Lewis [10] for a wide class of infinite-dimensional linear programming problems.

Let ν be a feasible point concentrated on h_1, \dots, h_n with masses $\lambda_1, \dots, \lambda_n$. Let s_1, \dots, s_m be m searcher trajectories of the set $\text{constr}(\nu)$. Let A be the payoff matrix generated by $\{h_1, \dots, h_n\}$ and $\{s_1, \dots, s_m\}$.

If $\text{rank}(A) = n$, we stop since ν is an extreme atomic measure. If not, we consider the subspace L defined by $L = \{\alpha \in R^n: A\alpha = 0\}$. Let $\hat{\alpha} = p_L(-e)$ be the orthogonal projection of $-e$ onto L . If $\hat{\sigma} = 0$, pick a nonzero $\hat{\alpha}$ arbitrarily in L .

We consider the atomic measure ν_ε defined by

$$\nu_\varepsilon = \sum_{i=1}^n (\lambda_i + \varepsilon \hat{\alpha}_i) \delta_{h_i}.$$

We wish to choose the greatest ε for which ν_ε is feasible, so we are moving as far as possible in the direction determined by $\hat{\alpha}$. The associated slack variable z_ε satisfies

$$z_\varepsilon(s) = z(s) + \varepsilon \sum_{i=1}^n \hat{\alpha}_i K(s, h_i) \quad \text{for all } s \in TS,$$

where z is the slack variable associated with ν . In particular, for the search trajectories s_j , we have $z_\varepsilon(s_j) = 0$, for $j = 1, \dots, m$, so that the algorithm moves in such a way as to maintain all the previous zeros of the slack variable. We define

$$\varepsilon_1 = \min \left\{ \frac{-\lambda_i}{\hat{\alpha}_i}: \hat{\alpha}_i < 0, i = 1, \dots, n \right\}.$$

Since the elements of the matrix A are positive, every nonzero element of L has a negative component, and therefore ε_1 is well defined. Thus $\lambda_i + \varepsilon \hat{\alpha}_i \geq 0$, for all $\varepsilon: 0 \leq \varepsilon \leq \varepsilon_1$ and $i = 1, \dots, n$.

We define

$$\varepsilon_2 = \min \left\{ \frac{-z(s)}{\sum_{i=1}^n \hat{\alpha}_i K(s, h_i)} \right\},$$

where the minimum is taken over $s \in S(O_s; h_1, \dots, h_n)$ such that the denominator is less than zero, and if there is no such s , then we take $\varepsilon_2 = +\infty$. We have $z_\varepsilon(s) \geq 0$, for all $s \in S(O_s; h_1, \dots, h_n)$, for all $0 \leq \varepsilon \leq \varepsilon_2$. Since the minimum of the slack variables z_ε is reached at an element of $S(O_s; h_1, \dots, h_n)$, ν is a feasible point for all ε such that $0 \leq \varepsilon \leq \varepsilon_0$ with $\varepsilon_0 = \min\{\varepsilon_1, \varepsilon_2\}$.

The result of this purification step is a value of the objective no greater than it was before, since the value of the objective functional at ν_ε is

$$\int_{TH} d\nu_\varepsilon(h) = \int_{TH} d\nu(h) + \varepsilon e^T \hat{\alpha}.$$

If $p_L(-e) = 0$, then $e^T \alpha = 0$ for all $\alpha \in L$, and therefore

$$\int_{TH} d\nu_\varepsilon(h) = \int_{TH} d\nu(h).$$

If $p_L(-e) \neq 0$, then $e + \hat{\alpha} \in L^\perp$, and we have

$$\begin{aligned} \int_{TH} d\nu_\varepsilon(h) &= \int_{TH} d\nu(h) + \varepsilon(e^T + \hat{\alpha}^T)\hat{\alpha} - \varepsilon \|\hat{\alpha}\|^2 \\ &= \int_{TH} d\nu(h) - \varepsilon \|\hat{\alpha}\|^2 < \int_{TH} d\nu(h). \end{aligned}$$

If after this step $\varepsilon = \varepsilon_1$, we repeat the entire process, starting from the feasible atomic measure ν_{ε_0} (deleting the h_i whose mass is zero) and the searcher trajectories s_1, \dots, s_m . If, on the other hand, $\varepsilon_0 = \varepsilon_2 < \varepsilon_1$, we repeat the entire process, starting from ν_{ε_0} and the searcher trajectories s_1, \dots, s_m, s^* , where s^* is a searcher trajectory that achieves the minimum in the definition of ε_2 . We prove below that by applying this process a finite number of times, we obtain an extreme atomic measure.

THEOREM 6.1. *The above purification algorithm terminates at an extreme atomic measure $\hat{\nu}$ in a finite number of iterations. Moreover, the value of the objective functional at $\hat{\nu}$ is no greater than the value at the initial feasible point.*

Proof. Every time $\varepsilon_0 = \varepsilon_1$, we drop at least one of the points h_i on which the measures are concentrated. This can happen, at most, n times. On the other hand, when $\varepsilon_0 = \varepsilon_2 < \varepsilon_1$, we introduce a searcher trajectory s^* , and thus the matrix A becomes the matrix A^* defined by

$$A^* = \begin{pmatrix} K(s_1, h_1) & \cdots & K(s_1, h_n) \\ \vdots & \ddots & \vdots \\ K(s_m, h_1) & \cdots & K(s_m, h_n) \\ K(s^*, h_1) & \cdots & K(s^*, h_n) \end{pmatrix}.$$

If $\text{rank}(A^*) \leq \text{rank}(A)$, then the last row is a linear combination of the other m rows. Since $\hat{\alpha} \in L$, $\sum_{i=1}^n \hat{\alpha}_i K(s_j, h_i) = 0$ for $j = 1, \dots, m$, and so $\sum_{i=1}^n \hat{\alpha}_i K(s^*, h_i) = 0$, in contradiction of the choice of s^* . Thus $\text{rank}(A^*) > \text{rank}(A)$ and in a finite number of steps, we obtain a matrix with full rank. \square

7. An optimality check. In this section we consider the problem of checking whether an extreme atomic measure of (HLP2) is optimal. For this we use the notion of reduced cost developed by Anderson and Nash [4].

Let ν be an extreme point concentrated on h_1, \dots, h_n with masses $\lambda_1, \dots, \lambda_n$. Since every $(\mu, y) \in B(\nu) \times B(z) \oplus N(A)$ can be expressed uniquely as $(\mu, y) = (\mu_B, y_B) + (\mu - \mu_B, y - y_B)$ with $(\mu_B, y_B) \in B(\nu) \times B(z)$ and $(\mu - \mu_B, y - y_B) \in N(A)$, the reduced-cost functional for ν is the map $\nu^*: B(\nu) \oplus N(A) \rightarrow R$ defined by

$$\langle (\mu, y), \nu^* \rangle = \int_{TH} d(\mu - \mu_B)(h).$$

It follows from the definition of $B(\nu)$ and $N(A)$ that μ_B is an atomic measure concentrated on h_1, \dots, h_n with masses $\alpha_1, \dots, \alpha_n$ and

$$y(s) - y_B(s) = \int_{TH} K(s, h) d\mu(h) - \sum_{i=1}^n \alpha_i K(s, h_i), \quad \text{for all } s \in TS.$$

Thus, since $y_B(s) = 0$ for all $s \in \text{constr}(\nu)$, we have

$$(2) \quad \sum_{i=1}^n \alpha_i K(s, h_i) = \int_{TH} K(s, h) d\mu(h) - y(s), \quad \text{for all } s \in \text{constr}(\nu).$$

Consider n searcher trajectories s_1, \dots, s_n of the set $\text{constr}(\nu)$ such that the payoff matrix A generated by them and $\{h_1, \dots, h_n\}$ has rank n . From (2) it follows that

$$(3) \quad \sum_{i=1}^n \alpha_i K(s_j, h_i) = \int_{TH} K(s_j, h) d\mu(h) - y(s_j), \quad \text{for } j = 1, \dots, n.$$

Therefore, defining

$$\hat{\mu}^T = \left(\int_{TH} K(s_1, h) d\mu(h), \dots, \int_{TH} K(s_n, h) d\mu(h) \right),$$

$$\hat{y}^T = (y(s_1), \dots, y(s_n)),$$

we can express $\alpha^T = (\alpha_1, \dots, \alpha_n)$ from (3) as $\alpha = A^{-1}(\hat{\mu} - \hat{y})$. Thus, if $\rho^T = e^T A^{-1}$, the reduced-cost functional ν^* satisfies

$$\begin{aligned} \langle (\mu, y), \nu^* \rangle &= \int_{TH} d\mu(h) - e^T \alpha \\ &= \int_{TH} d\mu(h) - \rho^T \hat{\mu} + \rho^T \hat{y} \\ &= \int_{TH} \left(1 - \sum_{j=1}^n \rho_j K(s_j, h) \right) d\mu(h) + \rho^T \hat{y}. \end{aligned}$$

This expression for the reduced-cost functional ν^* provides us with a check on the optimality of the extreme atomic measure ν . Essentially, the expression $1 - \sum_{j=1}^n \rho_j K(s_j, h)$ shows the effect on the objective function of introducing a particular new hider strategy h . If $\sum_{j=1}^n \rho_j K(s_j, h)$ is greater than 1, then an improvement can be obtained by introducing h into the current solution.

THEOREM 7.1. *Let ν be an extreme point concentrated on h_1, \dots, h_n with masses $\lambda_1, \dots, \lambda_n$. Let s_1, \dots, s_n be n searcher trajectories of the set $\text{constr}(\nu)$ such that the payoff matrix A generated by $\{h_1, \dots, h_n\}$ and $\{s_1, \dots, s_n\}$ has rank n . Let $\rho^T = e^T A^{-1}$. If (i) $\rho \geq 0$, (ii) $1 - \sum_{j=1}^n \rho_j K(s_j, h) \geq 0$ for all $h \in TH$, then ν is an optimal solution for (HLP1). Furthermore, the atomic measure μ concentrated on s_1, \dots, s_n with masses ρ_1, \dots, ρ_n is an optimal solution for (SLP1).*

Proof. The expressions (i) and (ii) imply that the atomic measure μ concentrated on s_1, \dots, s_n with masses ρ_1, \dots, ρ_n is a feasible point for (SLP1). Moreover, from Theorem 5.3 we know that $\lambda = A^{-1}e$, and therefore the value of the objective functional at ν satisfies

$$\int_{TH} d\nu(h) = e^T \lambda = e^T A^{-1} e = \rho^T e = \int_{TS} d\mu(s).$$

Using the weak duality theorem, the result follows. \square

The two conditions of the above theorem suggest an algorithm for (HLP2), which we describe in the next section.

8. An algorithm for the search game with mobile hider. The algorithm operates on the hider linear programming problem associated with the search game. It starts from an extreme atomic measure and tests for its optimality using the above optimality check. If it is not optimal, then an extreme atomic measure is produced that has a better (or, at least, no worse) value for the objective functional. This entire process is repeated until an optimal atomic measure is produced that satisfies the optimality check.

We begin by defining an initial extreme atomic measure. Let h_1 be any trajectory of the set TH . Let s_1 be the searcher trajectory in the set $S(O_s; h_1)$ described in § 4. Then the atomic measure $\nu = (1/K(s_1, h_1))\delta_{h_1}$ is a feasible solution to (HLP2). Moreover, the matrix generated by h_1 and s_1 is trivially of full rank; therefore ν is an extreme atomic measure for (HLP2).

Suppose now that we have a current solution ν that is an extreme atomic measure concentrated on h_1, \dots, h_n with masses $\lambda_1, \dots, \lambda_n$. Let s_1, \dots, s_n be n searcher trajectories of the set $\text{constr}(\nu)$. Let A be the payoff matrix generated by $\{h_1, \dots, h_n\}$ and $\{s_1, \dots, s_n\}$ of rank n .

As was proved in Theorem 5.3, the vector λ of the masses of ν satisfies the relation $\lambda = A^{-1}e$. Then, defining $\rho^T = e^T A^{-1}$ and denoting by Ω the value of the objective functional at the measure ν , it follows that $\Omega = e^T \lambda = \rho^T e$.

We can summarize the above information in a convenient tableau form as follows:

	h_1	\dots	h_n	
s_1	$K(s_1, h_1)$	\dots	$K(s_1, h_n)$	ρ_1
\vdots	\vdots	\ddots	\vdots	\vdots
s_n	$K(s_n, h_1)$	\dots	$K(s_n, h_n)$	ρ_n
	λ_1	\dots	λ_n	Ω

We say that $\{h_1, \dots, h_n\}$ is the hider basis and $\{s_1, \dots, s_n\}$ is the searcher basis.

Now, the optimality check of Theorem 7.1 applied to the extreme atomic measure ν yields three different cases, which we deal with in turn. Define $K(h)^T = (K(s_1, h), \dots, K(s_n, h))$.

Case 1. $\rho \geq 0$ and $\rho^T K(h) \leq 1$ for all $h \in TH$. In this case, the extreme atomic measure ν is an optimal solution to (HLP1). Furthermore, the atomic measure $\mu = \sum_{i=1}^n \rho_i \delta_{s_i}$ is an optimal solution to (SLP1), and both problems have the common value Ω .

Thus $1/\Omega$ is the value of the original search game, $\hat{\nu} = \sum_{i=1}^n (\lambda_i/\Omega) \delta_{h_i}$ is an optimal strategy for the hider, and $\hat{\mu} = \sum_{j=1}^n (\rho_j/\Omega) \delta_{s_j}$ is an optimal strategy for the searcher.

If either of the two conditions of Case 1 does not hold, it will indicate feasible directions of perturbation, which will produce an improvement in the value of the objective function.

Case 2. $\rho \not\geq 0$. Let k be an index such that $\rho_k < 0$. We consider the atomic measure ν_ε defined by $\nu_\varepsilon = \lambda(\varepsilon)_1 \delta_{h_1} + \dots + \lambda(\varepsilon)_n \delta_{h_n}$, where $\lambda(\varepsilon)$ is given by $\lambda(\varepsilon) = A^{-1}(e + \varepsilon e_k)$, with $e_k^T = (0, \dots, 1, \dots, 0)$. The slack variable z_ε associated with ν_ε satisfies

$$\begin{aligned} z_\varepsilon(s) &= K(s)^T \lambda(\varepsilon) - 1 \\ &= z(s) + \varepsilon K(s)^T A^{-1} e_k, \quad \text{for all } s \in TS, \end{aligned}$$

where $K(s)^T = (K(s, h_1), \dots, K(s, h_n))$. In particular, for the elements of the searcher basis s_j , we have

$$\begin{aligned} z_\varepsilon(s_j) &= 0, \quad \text{for } j = 1, \dots, n, \quad j \neq k, \\ z_\varepsilon(s_k) &= \varepsilon. \end{aligned}$$

We choose the largest ε that makes ν_ε feasible. This can be thought of as increasing the slack variable at the trajectory s_k while keeping the other trajectories s_j inside $\text{constr}(\nu_\varepsilon)$. The trajectory s_k leaves the searcher basis.

Now, since $\rho_k = e^T A^{-1} e_k < 0$, at least one of the components of $A^{-1} e_k$ is negative, and we may define

$$(4) \quad \varepsilon_1 = \min \left\{ \frac{-\lambda_i}{(A^{-1} e_k)_i} : (A^{-1} e_k)_i < 0, \quad \text{for } i = 1, \dots, n \right\}.$$

Thus $0 \leq \varepsilon_1 < +\infty$, and $\lambda(\varepsilon) \geq 0$, for all $\varepsilon: 0 \leq \varepsilon \leq \varepsilon_1$. Let

$$\varepsilon_2 = \min \left\{ \frac{-z(s)}{K(s)^T A^{-1} e_k} \right\},$$

where the minimum is taken over $s \in S(O_s; h_1, \dots, h_n)$ such that the denominator is less than zero, and, if there is no such s , then we take $\varepsilon_2 = +\infty$. We have

$$z_\varepsilon(s) \geq 0, \quad \text{for all } s \in S(O_s; h_1, \dots, h_n), \quad \text{for all } \varepsilon: 0 \leq \varepsilon \leq \varepsilon_2.$$

So, since the minimum of the slack variable z_ε on TS is reached at an element of $S(O_s; h_1, \dots, h_n)$, ν_ε is a feasible point for all ε such that $0 \leq \varepsilon \leq \varepsilon_0$ with $\varepsilon_0 = \min \{\varepsilon_1, \varepsilon_2\}$.

The value of the objective functional at ν_ε is

$$\begin{aligned} \int_{TH} d\nu_\varepsilon(h) &= e^T \lambda(\varepsilon) \\ &= e^T \lambda + \varepsilon e^T A^{-1} e_k \\ &= \int_{TH} d\nu(h) + \varepsilon \rho_k. \end{aligned}$$

Thus we obtain an improvement by moving from ν to ν_{ε_0} . From the form of this expression, it is natural to choose the index k such that ρ_k is the smallest component of ρ . Without loss of generality, we assume that $k = n$ to facilitate the notation.

Case 2.1. $\varepsilon_0 = \varepsilon_1$. In this case, the move we make is blocked because of the restriction on nonnegativity of the masses. We delete from the hider basis a trajectory whose mass becomes zero or, in other words, a trajectory where the minimum in (4) is attained. Let h_1 be such a trajectory. Then we have the following incomplete tableau:

	h_2	\dots	h_n	
s_1	$K(s_1, h_2)$	\dots	$K(s_1, h_n)$	
\vdots	\vdots	\ddots	\vdots	
s_{n-1}	$K(s_{n-1}, h_2)$	\dots	$K(s_{n-1}, h_n)$	
	$\lambda(\varepsilon_0)_2$	\dots	$\lambda(\varepsilon_0)_n$	$\Omega + \varepsilon_0 \rho_n$

If the payoff matrix generated by the hider trajectories $\{h_2, \dots, h_n\}$ and the searcher trajectories $\{s_1, \dots, s_{n-1}\}$ has rank $n-1$, then ν_{ε_0} is an extreme atomic measure. If not, we apply the purification algorithm described in § 6, which yields an extreme atomic measure with associated matrix of full rank. We repeat the entire process of the algorithm.

Case 2.2. $\varepsilon_0 = \varepsilon_2 < \varepsilon_1$. In this case, the move we make is blocked because of the restriction on nonnegativity of the slack variable z_ε . In other words, there is a searcher trajectory $s^* \in S(O_s; h_1, \dots, h_n)$ with $K(s^*) A^{-1} e_k < 0$, such that the value of the slack variable z_ε at s^* becomes zero when the value of ε_2 is reached. The searcher trajectory s^* enters the searcher basis and we have the following incomplete tableau:

	h_1	\dots	h_n	
s_1	$K(s_1, h_1)$	\dots	$K(s_1, h_n)$	
\vdots	\vdots	\ddots	\vdots	
s_{n-1}	$K(s_{n-1}, h_1)$	\dots	$K(s_{n-1}, h_n)$	
s^*	$K(s^*, h_1)$	\dots	$K(s^*, h_n)$	
	$\lambda(\varepsilon_0)_1$	\dots	$\lambda(\varepsilon_0)_n$	$\Omega + \varepsilon_0 \rho_n$

Now, ν_{ε_0} is an extreme atomic measure, since if the payoff matrix generated by $\{h_1, \dots, h_n\}$ and $\{s_1, \dots, s_{n-1}, s^*\}$ did not have rank n , then the n th row would be a linear combination of the other rows, and so $K(s^*)A^{-1}e^k = 0$. Hence we may repeat the entire process of the algorithm.

Case 3. $\rho \geq 0$ but $\rho^T K(h) \not\leq 1$ for all $h \in TH$. Let h_{n+1} be an element of TH such that $\rho^T K(h_{n+1}) = \sum_{i=1}^n \rho_i K(s_i, h_{n+1}) > 1$. We consider the atomic measure

$$\nu_\varepsilon = \lambda(\varepsilon)_1 \delta_{h_1} + \dots + \lambda(\varepsilon)_n \delta_{h_n} + \varepsilon \delta_{h_{n+1}},$$

where $\lambda(\varepsilon)$ is given by $\lambda(\varepsilon) = A^{-1}(e - \varepsilon K(h_{n+1}))$. The slack variable z_ε associated with ν_ε satisfies

$$\begin{aligned} z_\varepsilon(s) &= K(s)^T \lambda(\varepsilon) + \varepsilon K(s, h_{n+1}) - 1 \\ &= z(s) + \varepsilon(K(s, h_{n+1}) - K(s)^T A^{-1} K(h_{n+1})), \quad \text{for all } s \in TS. \end{aligned}$$

In particular, for the elements of the searcher basis, we have

$$\begin{aligned} z_\varepsilon(s_j) &= K(s_j)^T A^{-1} e - \varepsilon K(s_j)^T A^{-1} K(h_{n+1}) + \varepsilon K(s_j, h_{n+1}) - 1 \\ &= 1 - \varepsilon K(s_j, h_{n+1}) + \varepsilon K(s_j, h_{n+1}) - 1 \\ &= 0, \quad j = 1, \dots, n. \end{aligned}$$

The solution ν_ε may be thought of as increasing the mass of the pure hider strategy h_{n+1} while keeping the searcher trajectories s_j inside $\text{constr}(\nu_\varepsilon)$. The point h_{n+1} enters the hider basis. As before, we wish to choose the largest ε for which ν_ε is feasible. By hypothesis, $e^T A^{-1} K(h_{n+1}) > 1$, therefore at least one of the components of $A^{-1} K(h_{n+1})$ is positive and we may define the following value:

$$(5) \quad \varepsilon_1 = \min \left\{ \frac{\lambda_i}{(A^{-1} K(h_{n+1}))_i} : (A^{-1} K(h_{n+1}))_i > 0, i = 1, \dots, n \right\}.$$

Thus $0 \leq \varepsilon_1 < +\infty$, and $\lambda(\varepsilon) \geq 0$, for all $\varepsilon: 0 \leq \varepsilon \leq \varepsilon_1$.

We define the value ε_2 as follows:

$$\varepsilon_2 = \min \left\{ \frac{-z(s)}{K(s, h_{n+1}) - K(s)^T A^{-1} K(h_{n+1})} \right\},$$

where the minimum is taken over $s \in S(O_s; h_1, \dots, h_{n+1})$ such that the denominator is less than zero, and if there is no such s , then we take the value $\varepsilon_2 = +\infty$. We have

$$z_\varepsilon(s) \geq 0, \quad \text{for all } s \in S(O_s; h_1, \dots, h_{n+1}), \quad \text{for all } \varepsilon: 0 \leq \varepsilon \leq \varepsilon_2.$$

Thus, since the minimum of the slack variable z_ε on TS is reached at an element of $S(O_s; h_1, \dots, h_n)$, ν_ε is a feasible solution for all $\varepsilon: 0 \leq \varepsilon \leq \varepsilon_0$ with $\varepsilon_0 = \min\{\varepsilon_1, \varepsilon_2\}$.

The value of the objective functional at ν_ε is

$$\begin{aligned} \int_{TH} d\nu_\varepsilon(h) &= e^T \lambda(\varepsilon) + \varepsilon \\ &= e^T \lambda - \varepsilon e^T A^{-1} K(h_{n+1}) + \varepsilon \\ &= \int_{TH} d\nu(h) + \varepsilon(1 - \rho^T K(h_{n+1})). \end{aligned}$$

Thus we obtain an improvement by moving from ν to ν_{ε_0} .

In this case, there is no obvious criterion for the choice of h_{n+1} . In practice, we take h_{n+1} to be any element of TH such that $1 < \rho^T K(h_{n+1})$.

Case 3.1. $\varepsilon_0 = \varepsilon_1$. In this case, the move we make is blocked because of the restriction on nonnegativity of the masses. We delete from the hider basis a trajectory whose mass becomes zero, or, in other words, a trajectory where the minimum in (5) is attained. Let h_1 be such a trajectory. Then if we write $\Omega_0 = \Omega + \varepsilon(1 - \rho^T K(h_{n+1}))$ we have the following incomplete tableau:

	h_2	\cdots	h_n	h_{n+1}	
s_1	$K(s_1, h_2)$	\cdots	$K(s_1, h_n)$	$K(s_1, h_{n+1})$	
\vdots	\vdots	\ddots	\vdots	\vdots	
s_n	$K(s_n, h_2)$	\cdots	$K(s_n, h_n)$	$K(s_n, h_{n+1})$	
	$\lambda(\varepsilon_0)_2$	\cdots	$\lambda(\varepsilon_0)_n$	ε_0	
					Ω_0

If the payoff matrix generated by the hider trajectories $\{h_2, \cdots, h_{n+1}\}$ and the searcher trajectories $\{s_1, \cdots, s_n\}$ has rank $n-1$, then ν_{ε_0} is an extreme atomic measure. If not, we apply the purification algorithm, which yields an extreme atomic measure with associated matrix of full rank. We repeat the entire process of the algorithm.

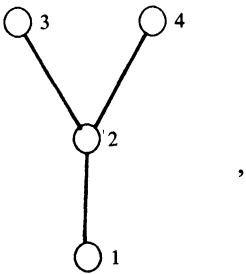
Case 3.2. $\varepsilon_0 = \varepsilon_2 < \varepsilon_1$. In this case, the move we make is blocked because of the restriction on nonnegativity of the slack variable z_{ε} . In other words, there is a searcher trajectory $s^* \in S(O_s; h_1, \cdots, h_{n+1})$ with $K(s^*, h_{n+1}) - K(s^*)A^{-1}K(h_{n+1}) < 0$, where the value ε_2 is reached and therefore the slack variable z_{ε_0} becomes zero. This searcher trajectory s^* enters the searcher basis and we have the following incomplete tableau:

	h_1	\cdots	h_n	h_{n+1}	
s_1	$K(s_1, h_1)$	\cdots	$K(s_1, h_n)$	$K(s_1, h_{n+1})$	
\vdots	\vdots	\ddots	\vdots	\vdots	
s_n	$K(s_n, h_1)$	\cdots	$K(s_n, h_n)$	$K(s_n, h_{n+1})$	
s^*	$K(s^*, h_1)$	\cdots	$K(s^*, h_n)$	$K(s^*, h_{n+1})$	
	$\lambda(\varepsilon_0)_1$	\cdots	$\lambda(\varepsilon_0)_n$	ε_0	Ω_0

Now, ν_{ε_0} is an extreme atomic measure, since if the payoff matrix generated by $\{h_1, \cdots, h_{n+1}\}$ and $\{s_1, \cdots, s_n, s^*\}$ did not have rank $n+1$, then the last row would be a linear combination of the other rows, and so $K(s^*, h_{n+1}) - K(s^*)^T A^{-1} K(h_{n+1}) = 0$. Therefore we may repeat the entire process of the algorithm.

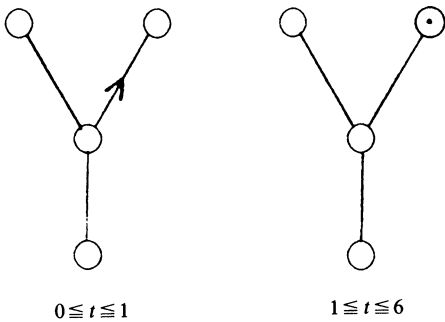
Summing up, the algorithm produces a sequence of extreme atomic measures to (HLP1) with nonincreasing values for the objective functional until the optimality check is satisfied. When this occurs, we obtain an atomic probability measure that is an optimal strategy for the hider, an atomic probability measure that is an optimal strategy for the searcher, and the value for the game. The algorithm works entirely with atomic measures. Therefore we cannot hope for finite convergence when the network is such that the optimal solution is not atomic (as will be the case for the majority of networks). Nevertheless, it may be that the result of applying the steps of the algorithm repeatedly is an atomic measure that approaches optimality. Whether such a convergence property can be established for our algorithm remains an open question. Certainly, the proof of such a result would require us to more closely specify the hider strategy to be introduced in Case 3 of the algorithm.

9. An example. Let Q be the network defined as follows:



where each arc is of length 1. Let node 1 be the starting point for the searcher, node 2 be the starting point for the hider, and 6 be the termination time.

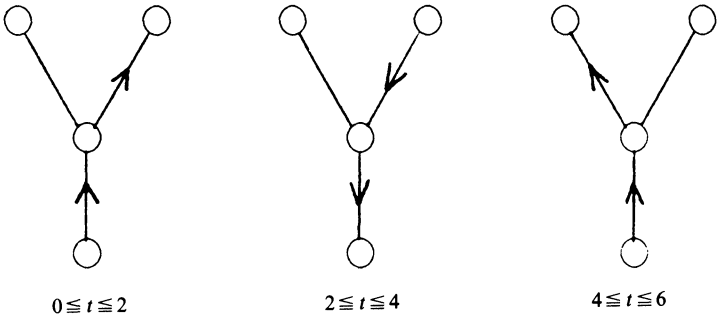
Iteration 1. Let h_1 be the following hider trajectory:



We indicate movement along the network by arrows. The dot at node 4 means that the hider remains immobile there for the period of time $[1], [6]$. Then $\nu_1 = \frac{1}{2}\delta_{h_1}$ is an initial extreme atomic measure for the algorithm. The first tableau is

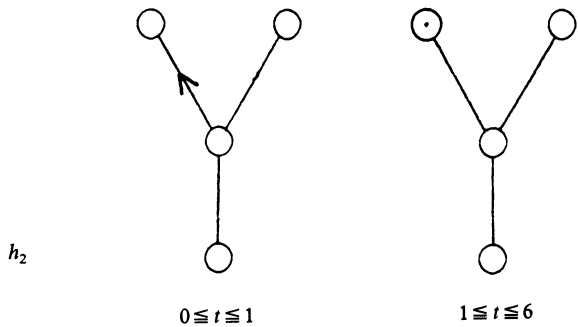
	h_1	
s_1	2	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{1}{2}$

where s_1 is the searcher trajectory defined as follows:

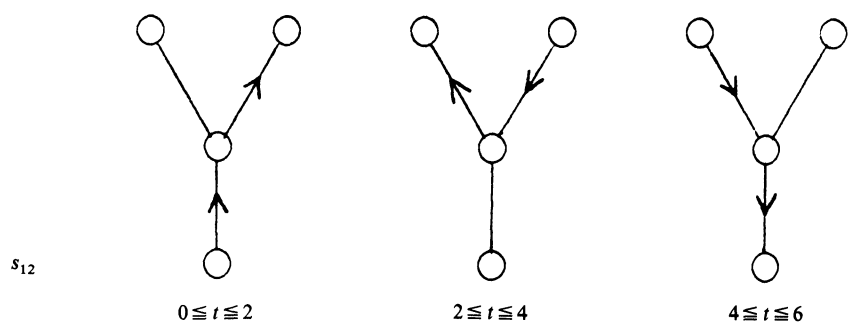


The extreme atomic measure $\nu_1 = \frac{1}{2}\delta_{h_1}$ satisfies Case 3: $\rho \geq 0$ but $\rho^T K(h) \not\leq 1$ for all

$h \in TH$. Let h_2 be the following hider trajectory:



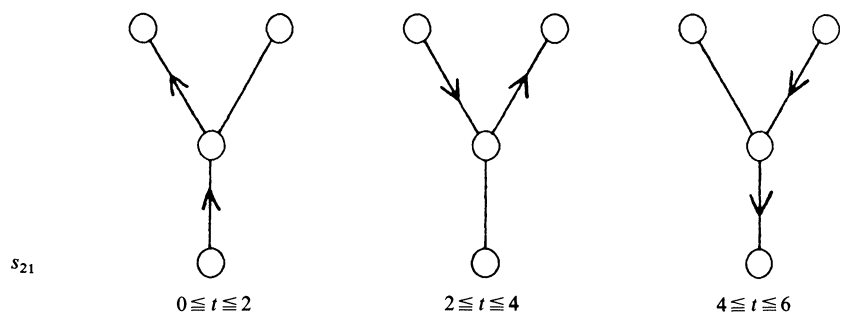
Then $\rho^T K(h_2) = 3$ and h_2 is introduced into the hider basis. Then $\varepsilon_0 = \varepsilon_2 = 0$, which is reached at the searcher trajectory s_{12} , defined as follows:



We obtain the extreme atomic measure $\nu_2 = \frac{1}{2}\delta_{h_1} + 0\delta_{h_2}$ with the matrix generated by $\{h_1, h_2\}$ and $\{s_1, s_{12}\}$ of full rank. Its associated tableau is

	h_1	h_2	
s_1	2	6	$-\frac{1}{2}$
s_{12}	2	4	1
	$\frac{1}{2}$	0	$\frac{1}{2}$

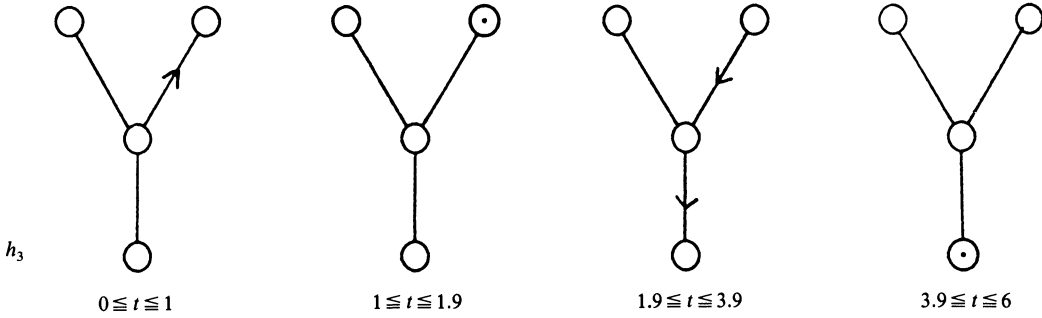
Iteration 2. The extreme atomic measure $\nu_2 = \frac{1}{2}\delta_{h_1} + 0\delta_{h_2}$ satisfies Case 2: $\rho \not\geq 0$. The trajectory s_1 leaves the hider basis. Then $\varepsilon_0 = \varepsilon_2 = \frac{1}{2}$, which is reached at the searcher trajectory s_{21} , defined as follows:



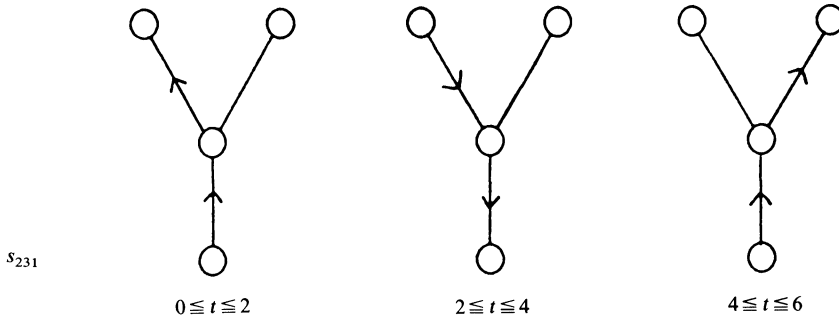
We obtain the extreme atomic measure $\nu_3 = \frac{1}{6}\delta_{h_1} + \frac{1}{6}\delta_{h_2}$ with the matrix generated by $\{h_1, h_2\}$ and $\{s_{21}, s_{12}\}$ of full rank. Its associated tableau is

	h_1	h_2	
s_{21}	4	2	$\frac{1}{6}$
s_{12}	2	4	$\frac{1}{6}$
	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{6}$

Iteration 3. The extreme atomic measure $\nu_3 = \frac{1}{6}\delta_{h_1} + \frac{1}{6}\delta_{h_2}$ satisfies Case 3: $\rho \geq 0$ but $\rho^T K(h) \not\leq 1$ for all $h \in TH$. Let h_3 be the following hider trajectory:



Then $\rho^T K(h_3) = 1.3249$, and h_3 is introduced into the hider basis. In this step $\varepsilon_0 = \varepsilon_2 = 0.0623$, which is reached at the searcher trajectory s_{231} , defined as follows:

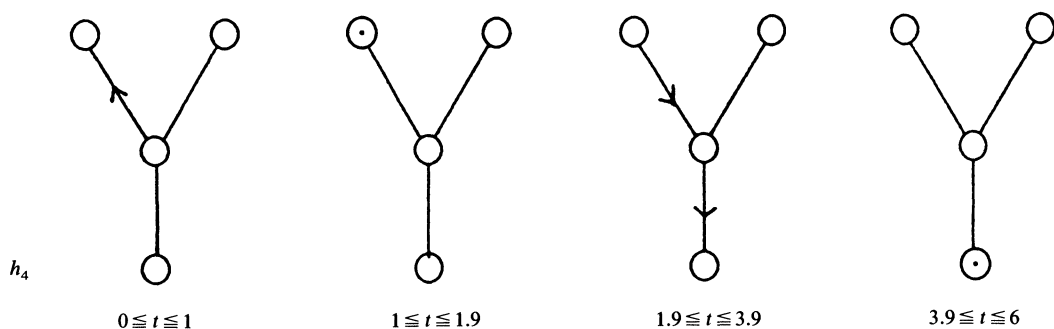


We obtain the extreme atomic measure $\nu_4 = 0.0623\delta_{h_1} + 0.1884\delta_{h_2} + 0.0623\delta_{h_3}$ with the matrix generated by $\{h_1, h_2, h_3\}$ and $\{s_{21}, s_{12}, s_{231}\}$ of full rank. Its associated tableau is

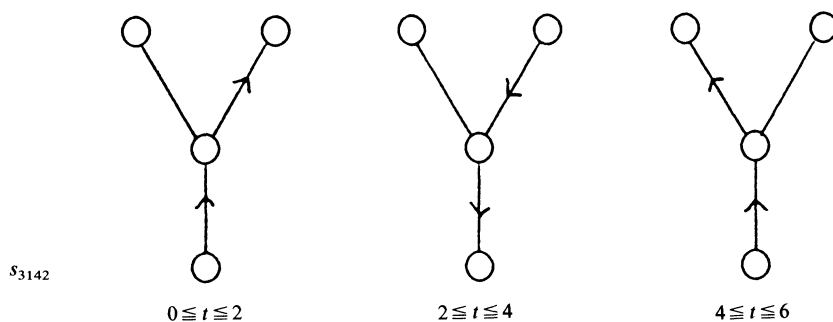
	h_1	h_2	h_3	
s_{21}	4	2	6	0.0654
s_{12}	2	4	1.95	0.1869
s_{231}	6	2	4	0.0607
	0.0623	0.1844	0.0623	0.3130

Iteration 4. The extreme atomic measure $\nu_4 = 0.0623\delta_{h_1} + 0.1884\delta_{h_2} + 0.0623\delta_{h_3}$ satisfies Case 3: $\rho \geq 0$ but $\rho^T K(h) \not\leq 1$ for all $h \in TH$. Let h_4 be the following hider

trajectory:



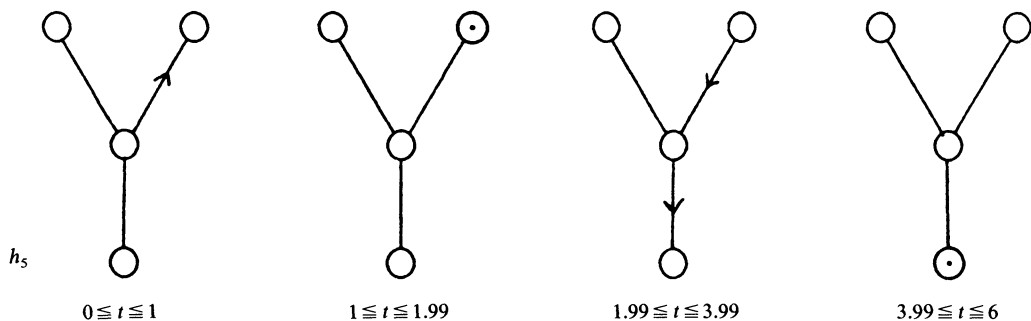
Then $\rho^T K(h_4) = 1.2518$, and h_4 is introduced into the hider basis. At this step $\varepsilon_0 = \varepsilon_2 = 0.0716$, which is reached at the searcher trajectory s_{3142} , defined as follows:



We obtain the extreme atomic measure $\nu_5 = 0.0716\delta_{h_1} + 0.0716\delta_{h_2} + 0.0716\delta_{h_3} + 0.0716\delta_{h_4}$ with the matrix generated by $\{h_1, h_2, h_3, h_4\}$ and $\{s_{21}, s_{12}, s_{231}, s_{3142}\}$ of full rank. Its associated tableau is

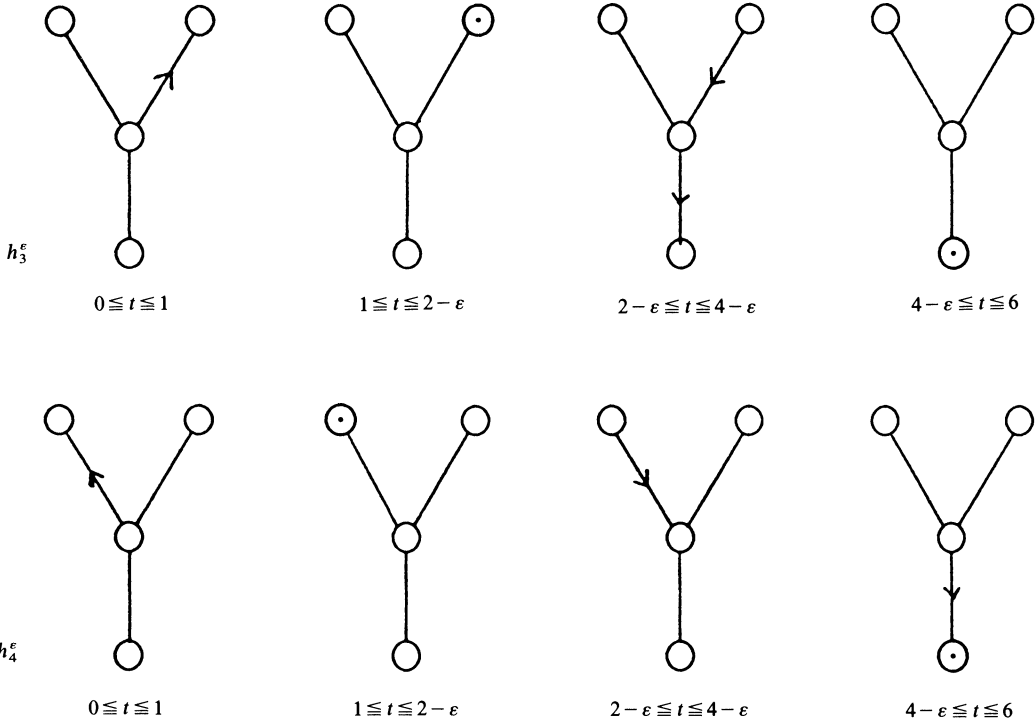
	h_1	h_2	h_3	h_4	
s_{21}	4	2	6	1.95	0.0734
s_{12}	2	4	1.95	6	0.0734
s_{231}	6	2	4	1.95	0.0698
s_{3142}	2	6	1.95	4	0.0698
	0.0716	0.0716	0.0716	0.0716	0.2867

Iteration 5. The extreme atomic measure $\nu_5 = 0.0716\delta_{h_1} + 0.0716\delta_{h_2} + 0.0716\delta_{h_3} + 0.0716\delta_{h_4}$ satisfies Case 3: $\rho \geq 0$ but $\rho^T K(h) \not\leq 1$ for all $h \in TH$. Let h_5 be the following hider trajectory:



Then $\rho^T K(h_5) = 1.0016$ and h_5 is introduced into the hider basis. At this step $\varepsilon_0 = \varepsilon_1 = 0.0724$, which is reached at the hider trajectory h_3 . This trajectory leaves the hider basis, and we obtain the extreme atomic measure $\nu_6 = 0.0708\delta_{h_1} + 0.0713\delta_{h_2} + 0.0713\delta_{h_4} + 0.0724\delta_{h_5}$.

At Iteration 6, a new hider trajectory h_6 is introduced. This is symmetrical with h_5 (in the same way that h_4 is symmetrical with h_3). We can carry on indefinitely, introducing new hider trajectories at each stage, loitering a little longer at nodes 3 and 4. Note that the trajectories of the searcher basis always remain as $\{s_{21}, s_{12}, s_{231}, s_{3142}\}$. Let $h_3^\varepsilon, h_4^\varepsilon$ be the following hider trajectories:



Thus in the fourth iteration, $\varepsilon = 0.1$, and in the sixth iteration, $\varepsilon = 0.01$. Taking limits as $\varepsilon \rightarrow 0$, the tableau becomes

	h_1	h_2	h_3^ε	h_4^ε	
s_{21}	4	2	6	2	$\frac{1}{14}$
s_{12}	2	4	2	6	$\frac{1}{14}$
s_{231}	6	2	4	2	$\frac{1}{14}$
s_{3142}	2	6	2	4	$\frac{1}{14}$
	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{4}{14}$

We now consider the searcher strategy $\hat{\mu}$ defined by

$$\hat{\mu} = \frac{1}{14} \delta_{s_{21}} + \frac{1}{14} \delta_{s_{12}} + \frac{1}{14} \delta_{s_{231}} + \frac{1}{14} \delta_{s_{3142}}.$$

It is easy to see that

$$\hat{\omega}(h) = 1 - \frac{1}{14} K(s_{21}, h) - \frac{1}{14} K(s_{12}, h) - \frac{1}{14} K(s_{231}, h) - \frac{1}{14} K(s_{3142}, h) \geq 0$$

for all $h \in TH$. Thus $\hat{\mu}$ is a feasible point for (SLP1) and the value of the objective functional at $\hat{\mu}$ is $4/14$.

On the other hand, we consider the hider strategy $\hat{\nu}_\varepsilon$ defined by

$$\hat{\nu}_\varepsilon = \frac{1}{14 - \varepsilon/2} \delta_{h_1} + \frac{1}{14 - \varepsilon/2} \delta_{h_2} + \frac{1}{14 - \varepsilon/2} \delta_{h_3^\varepsilon} + \frac{1}{14 - \varepsilon/2} \delta_{h_4^\varepsilon}.$$

It is easy to see that

$$\begin{aligned} \hat{z}(s) &= \frac{1}{14 - \varepsilon/2} K(s, h_1) + \frac{1}{14 - \varepsilon/2} K(s, h_2) + \frac{1}{14 - \varepsilon/2} K(s, h_3^\varepsilon) \\ &\quad + \frac{1}{14 - \varepsilon/2} K(s, h_4^\varepsilon) - 1 \geq 0 \end{aligned}$$

for all $s \in S(O_s; h_1, h_2, h_3^\varepsilon, h_4^\varepsilon)$. Thus $\hat{\nu}_\varepsilon$ is a feasible point of (HLP1) and the value of the objective functional at this point is $4/(14 - \frac{\varepsilon}{2})$. Applying the weak duality theorem, it follows that $\hat{\mu}$ is an optimal solution for (SLP1), $\hat{\nu}_\varepsilon$ is an ε -optimal solution for (HLP1), and the value of both programs is $4/14$.

Thus we have established that the value of the game is $7/2$, $\mu = \frac{1}{4}\delta_{s_{21}} + \frac{1}{4}\delta_{s_{12}} + \frac{1}{4}\delta_{s_{231}} + \frac{1}{4}\delta_{s_{3142}}$ is an optimal strategy for the searcher, and $\nu_\varepsilon = \frac{1}{4}\delta_{h_1} + \frac{1}{4}\delta_{h_2} + \frac{1}{4}\delta_{h_3^\varepsilon} + \frac{1}{4}\delta_{h_4^\varepsilon}$ is an ε -optimal strategy for the hider.

REFERENCES

- [1] S. ALPERN, *The search game with mobile hider on the circle*, in Differential Games and Control Theory, E. O. Roxin, P. Lin, and R. L. Sternberg, eds., Lecture Notes in Pure and Appl. Math. 10, Marcel Dekker, New York, 1974, pp. 181–200.
- [2] S. ALPERN AND M. ASIC, *The search value of a network*, Networks, 15 (1985), pp. 229–238.
- [3] ———, *Ambush strategies in search games on graphs*, SIAM J. Control Optim., 14 (1986), pp. 66–75.
- [4] E. J. ANDERSON AND P. NASH, *Linear Programming in Infinite-Dimensional Spaces*, John Wiley, Chichester, UK, 1987.
- [5] C. H. FITZGERALD, *The princess and monster differential game*, SIAM J. Control Optim., 17 (1979), pp. 700–712.
- [6] J. G. FOREMAN, *The princess and the monster on the circle*, in Differential Games and Control Theory, E. O. Roxin, P. Lin, and R. L. Sternberg, eds., Lecture Notes in Pure and Appl. Math. 10, Marcel Dekker, New York, 1974, pp. 231–240.
- [7] ———, *Differential search games with mobile hider*, SIAM J. Control Optim., 15 (1977), pp. 841–856.
- [8] S. GAL, *Search games*, Math. Sci. Engrg., Vol. 149, Academic Press, New York, 1980.
- [9] R. ISSAACS, *Differential Games*, John Wiley, New York, 1965.
- [10] A. S. LEWIS, *Extreme points and purification algorithms in general linear programming*, in Infinite Programming, E. J. Anderson and A. B. Philpott, eds., Lecture Notes in Econom. and Math. Systems, 259, Springer-Verlag, Berlin, 1985, pp. 123–135.
- [11] M. I. ZELIKIN, *On a differential game with incomplete information*, Soviet Math. Dokl., 13 (1972), pp. 228–230.

MINIMAL LENGTH CURVES THAT ARE NOT EMBEDDABLE IN AN OPEN PLANAR SET: THE PROBLEM OF A LOST SWIMMER WITH A COMPASS*

R. HASSIN[†] AND A. TAMIR^{† ‡}

Abstract. Given an open bounded set S in R^2 , the problem of computing a path f of minimum size such that for every $x \in S$ the set $\{x\} + f$ intersects the boundary of S is considered. The existence of such paths is proved both when the path size is its length and when it is its (one-dimensional Hausdorff outer) measure. Some theorems characterizing optimal paths are proved and it is shown that when S is convex, the minimum width chords of $\text{Cl}(S)$ are optimal with respect to both size definitions.

Key words. search theory, computational geometry, planar convex sets

AMS(MOS) subject classifications. 90B40, 52A10, 49A40

1. Introduction. A fisherman on a small boat lost on a big lake in a very thick fog has no information regarding his location. He has zero visibility but possesses a compass and a map of the lake and its surroundings. The fisherman can do dead-reckoning navigation by selecting at each point in time an azimuth and by traveling along this direction for any distance d he wishes to cover. His objective is to minimize the distance he must travel to the shoreline.

Next consider a soldier lost in a mine field under zero visibility conditions. He has a compass, a map of the field (which does not indicate the mines), as well as a special spoke to search for the mines. The soldier wants to minimize the time he will need to reach a boundary of the field. If he traverses a certain segment of a path for the first time, he must search for mines, thus moving at a very low speed v . However, if his path repeats a segment for a second time, he can speed up, attaining a velocity that is practically infinite in comparison to v .

Suppose that both the fisherman and the soldier are conservative and that they wish to minimize the maximum travel distance (time) over all possible initial locations.

We use several examples to demonstrate the difference between the two models (see Fig. 1). The optimal path of the fisherman is given by the minimal length path, while that of the soldier is depicted by the minimal measure path in the examples below. For comparison purposes, we normalize v , the speed of the soldier, to one unit when he explores “new avenues,” and we let him repeat a segment he has already traveled before with infinite velocity. With this assumption, the measure of a path does not exceed its length. (These two terms are properly defined in the next section.) In Fig. 1, examples (a), (c), (d), and (f), the optimal measure path is an arc; i.e., none of its points is visited more than once. Therefore, the length is equal to the measure. In example (b), the minimal length is 2, while the minimal measure is only $1 + \sqrt{3}/2$. In example (e) the difference between the measure and the length is ε .

*Received by the editors June 11, 1990; accepted for publication (in revised form) April 18, 1991.

[†]Beverly and Raymond Sackler Faculty of Exact Sciences, School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel.

[‡]Department of Statistics and Operations Research, New York University, New York, New York 10003.

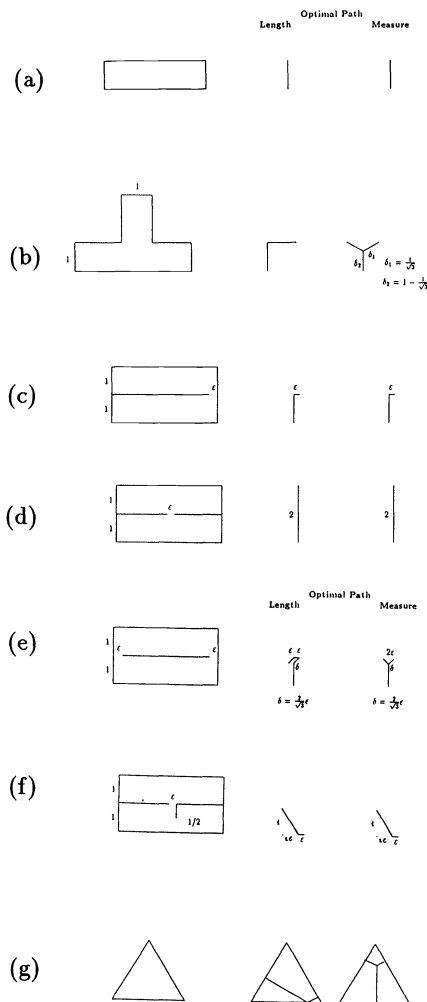


FIG. 1

Finally, example (d) demonstrates that the length and the measure are very sensitive to connectivity properties. In particular, they might both have discontinuity jumps when the boundary is slightly perturbed. The optimal length and measure are equal to 2 for any $\varepsilon > 0$. However, if $\varepsilon = 0$, i.e., the lake becomes disconnected, they both decrease to 1.

In this study, after we provide general existence theorems, we focus on convex open sets (lakes). We prove that the length and the measure are both equal to the width of the convex set, i.e., to the minimum distance between a pair of distinct parallel lines that bound the set. Therefore, an optimal path is a (shortest) line segment connecting this pair of lines. Example (g) demonstrates that the optimal path is not necessarily unique. It also shows that there may be nonlinear optimal paths even in the convex case. In particular, every path whose image is the three normals from a point in the equilateral triangle to its edges is a minimal measure path. A piecewise linear path with a single breakpoint on an edge of the equilateral triangle whose image is the two normals from that point to the other two edges is a minimal length path.

We do not know of any works dealing with optimal “navigation” with a compass. Works on navigation without a compass and related topics are described in [1], [3], and [5]–[14]. In the last section, we present a general framework unifying search problems of this nature.

2. The mathematical model. Let S be a bounded open set in R^2 . Let ∂S denote the boundary of S and let $\text{Cl}(S)$ denote its closure.

A *path* in R^2 is a continuous function f from the unit interval I in R^1 into R^2 . A path f is called an *arc* (a simple Jordan curve) if f is one-to-one. We use the symbol $f(I)$ to indicate the *image* of I , i.e., the set of all elements $f(t)$ in R^2 where $t \in I$.

The *path length* of f , $\lambda(f)$, is defined as

$$\lambda(f) = \lim_{n \rightarrow \infty} \sum_{k=0}^{n-1} \left\| f\left(\frac{k+1}{n}\right) - f\left(\frac{k}{n}\right) \right\| ,$$

where $\|\cdot\|$ is the Euclidean norm. When f is piecewise smooth (piecewise continuously differentiable), $\lambda(f) = \int_I \|f^{(1)}(t)\| dt$.

Let x, y be in R^2 and let A and B be subsets of R^2 . Define $d(x, y) = \|x - y\|$, $d(x, A) = d(A, x) = \inf \{d(x, y) \mid y \in A\}$, and $\text{diam}(A) = \sup \{d(x, y) \mid x, y \in A\}$. Let $d(A \triangle B)$ denote the Hausdorff distance between A and B , i.e.,

$$d(A \triangle B) = \max \left\{ \sup \{d(x, B) \mid x \in A\}, \sup \{d(y, A) \mid y \in B\} \right\} .$$

Also, define $A + B = \{x + y \mid x \in A, y \in B\}$.

Let C be a subset of R^2 . The *one-dimensional Hausdorff outer measure* of C in R^2 is

$$\lambda_1(C) = \lim_{\varepsilon \rightarrow 0} \left(\inf \left\{ \sum_{i=1}^{\infty} \text{diam}(A_i) \mid \bigcup_{i=1}^{\infty} A_i = C, \text{diam}(A_i) \leq \varepsilon \text{ for all } i \right\} \right) .$$

Consider a path f . Then f is *S-uncontained* if $f(0) = 0$ and for every $x \in S$ the set $\{x\} + f(I)$ intersects ∂S . The length of such a path f is given by $\lambda(f)$, defined above, and its *measure* is given by $\lambda_1(f(I))$.

Motivated by the examples presented above, we consider the following two optimization models.

Model 1. Find an *S-uncontained* path of minimum length.

Model 2. Find an *S-uncontained* path of minimum measure.

In Theorem 1, we prove the existence of optimal paths in either of the two models. It has been demonstrated above that the two models can have different solution paths. We show here that if the set S is convex, then both models are optimized by the same linear path. Specifically, an optimizer of both models is the *minimum width chord* of $\text{Cl}(S)$. (The latter is defined as a line segment in R^2 of minimum length connecting any two distinct parallel lines that bound $\text{Cl}(S)$ between them.)

THEOREM 1. *Let S be an open bounded set in R^2 .*

a) *There exists an S-uncontained path of minimum length.*

b) *There exists an S-uncontained path of minimum measure.*

Proof. Denote $M = \text{diam}(\text{Cl}(S))$. To prove the theorem, it is certainly sufficient to consider only those paths whose length (and measure) is bounded by M .

(a) Let β denote the infimum of the lengths of all *S-uncontained* paths. Consider an *S-uncontained* path f with length $\lambda(f)$. We represent f by a standard

parametrization \hat{f} on its length: $\hat{f} : I \rightarrow R^2$, where \hat{f} is continuous and $\hat{f}(0) = 0$. \hat{f} is M -Lipschitz. Using the Arzelà–Ascoli theorem [4, p. 266], we conclude that the set of all (parametrized) S -uncontained and M -Lipschitz paths is nonempty and compact in the uniform convergence topology. If $\{\hat{f}_n\}$ is a sequence of (parametrized) S -uncontained paths such that $\lambda(\hat{f}_n) \rightarrow \beta$, there exists a subsequence $\{\hat{f}_{n_i}\}$ converging uniformly to an S -uncontained path g . Since path length is lower semicontinuous, we obtain $\lambda(g) \leq \lim_i \lambda(\hat{f}_{n_i}) = \beta$.

(b) Let γ denote the infimum of the measures of all S -uncontained paths, and let $\{\hat{f}_n\}$ be a sequence of (parametrized) S -uncontained paths such that $\lambda_1(\hat{f}_n(I)) \rightarrow \gamma$. As in part (a), let $\{\hat{f}_{n_i}\}$ be a subsequence converging uniformly to an S -uncontained path h . It follows that the sets $\{\hat{f}_{n_i}(I)\}$ converge in Hausdorff distance to $h(I)$. Therefore, using [6, Thm. 3] we conclude that

$$\lambda_1(h(I)) \leq \lim_i \lambda_1(\hat{f}_{n_i}(I)) = \gamma. \quad \square$$

Remark. Instead of using the Arzelà–Ascoli theorem in the above proof, one might prefer a more elementary argument as exhibited in [3]. It is based on repeatedly using the Bolzano–Weierstrass property in the usual plane metric.

We will need the following definition: Let $\alpha \in I$. The *subpath* of f defined by α , f^α , is

$$f^\alpha(t) = \begin{cases} f(t), & 0 \leq t \leq \alpha \\ f(\alpha), & \alpha \leq t \leq 1. \end{cases}$$

THEOREM 2. *Let f be an S -uncontained path. Then there exist $\bar{x} \in \text{Cl}(S)$ and $\alpha \in I$ such that f^α is S -uncontained and $\{\bar{x}\} + f^\alpha(I) \subseteq \text{Cl}(S)$.*

Proof. For each $x \in S$ define

$$t(x) = \sup \{t \in I \mid x + f(s) \in S \text{ for all } 0 \leq s < t\}.$$

Since f is S -uncontained, $x + f(t(x))$ is in ∂S .

Define $\alpha = \sup \{t(x) \mid x \in S\}$.

For each $x \in S$, $\{x\} + f^\alpha(I)$ intersects ∂S . Thus, f^α is S -uncontained. From the definition of α and the compactness of $\text{Cl}(S)$, there exists a sequence $\{x^n\}$ of points in S that converges to some $\bar{x} \in \text{Cl}(S)$, and $\{t(x^n)\}$ converges to α . We claim that $\{\bar{x}\} + f^\alpha(I) \subseteq \text{Cl}(S)$.

Suppose, by contradiction, that there exists some s , $0 < s < \alpha$, and $\bar{x} + f^\alpha(s) \notin \text{Cl}(S)$. Let $\varepsilon = d(\bar{x} + f^\alpha(s), \text{Cl}(S)) > 0$. Let n be such that $d(x^n, \bar{x}) < \varepsilon/2$, and $s < t(x^n) \leq \alpha$. Then for any $y \in \text{Cl}(S)$, $d(y, x^n + f^\alpha(s)) \geq d(y, \bar{x} + f^\alpha(s)) - d(x^n, \bar{x}) \geq \varepsilon/2$. Therefore, $d(x^n + f^\alpha(s), \text{Cl}(S)) \geq \varepsilon/2$ for some $s < t(x^n)$. This contradicts the definition of $t(x^n)$. \square

Next we prove that if S is a bounded open convex set in R^2 , then there exists an S -uncontained path of minimum length and minimum measure that is a linear function. In particular, we show that the length of such a linear path is the width of $\text{Cl}(S)$.

We now need the following definition.

DEFINITION. Let u be a point in $\text{Cl}(S)$. Then d in R^2 is a *feasible direction* of S at u if there exists $\varepsilon > 0$ such that $u + \varepsilon d$ is in S . Otherwise d is called *infeasible*.

THEOREM 3. *Let f be an S -uncontained path for a bounded open convex set $S \subseteq R^2$, and let $x \in \text{Cl}(S)$ satisfy $\{x\} + f(I) \subseteq \text{Cl}(S)$. Define the tangency set $\hat{R}(f, x)$*

$$\hat{R}(f, x) = \partial S \cap (\{x\} + f(I)) .$$

Then one of the following holds:

- (1) *There exist two distinct and parallel supporting (subgradient) lines to $\text{Cl}(S)$ at some pair of points in $\hat{R}(f, x)$.*
- (2) *There exist three distinct supporting lines to $\text{Cl}(S)$ at some points in $\hat{R}(f, x)$, such that the triangle generated by these lines contains $\text{Cl}(S)$.*

Proof. Since f is S -uncontained, it follows that there is no direction d at x and an $\varepsilon > 0$ such that the set $\{x + \varepsilon d\} + f(I)$ is contained in S . The points in $\hat{R}(f, x)$ block any translation of the set $\{x\} + f(I)$ into S . Formally, it follows that the set of infeasible directions at all the points in $\hat{R}(f, x)$ exhausts all directions in R^2 .

Consider a supporting line ℓ to $\text{Cl}(S)$ at some point u in $\hat{R}(f, x)$. With each infeasible direction at u , we associate a point on the unit circle corresponding to its angle from the horizontal x_1 -axis. Thus the supporting line ℓ is associated with a closed subarc I_ℓ of the unit circle of length π . (I_ℓ captures all infeasible directions defined by the half plane not containing $\text{Cl}(S)$.)

Consider next the collection of subarcs obtained by looking at all points in $\hat{R}(f, x)$ and their supporting lines. From the above, it follows that the union of all the subarcs is the unit circle. If there exists a pair of subarcs whose union is the unit circle, then (1) holds. Otherwise, the union of the interiors of the subarcs is again the unit circle. Due to compactness of the unit circle, there is a finite subcollection of subarcs whose union is the unit circle. To summarize, there is a finite number of at least three supporting lines at points in $\hat{R}(f, x)$ that define a convex bounded polygon containing $\text{Cl}(S)$. Since no pair of these lines is parallel, it is a simple matter to verify inductively that any such polygon can be bounded by a triangle formed by three of its supporting lines. This completes the proof. \square

LEMMA 1. *Let $\{\ell_1, \ell_2, \ell_3\}$ be a collection of three distinct pairwise nonparallel lines in R^2 . Then the minimum over R^2 of the sum of (Euclidean) distances from the three lines is attained at a point where two of these lines intersect.*

Proof. For x in R^2 , let $g_i(x) = d(x, \ell_i)$, $i = 1, 2, 3$, and $g(x) = g_1(x) + g_2(x) + g_3(x)$.

Consider an arbitrary line L in R^2 . If L and ℓ_i are parallel, the restriction of $g_i(x)$ to L is a constant function. Otherwise, this restriction is piecewise linear with one breakpoint at the intersection point of L and ℓ_i . Therefore, the restriction of $g(x)$ to L is a (convex) piecewise linear function having at most three breakpoints (the intersection points of L with the three given lines). The minimum of $g(x)$ over L is attained at an intersection point of L with some line ℓ_i , $i = 1, 2, 3$.

Let x^* be a minimum point of $g(x)$ over R^2 , and consider some line L containing x^* . Then there exists some line ℓ_i and the point z^* , $\{z^*\} = L \cap \ell_i$, such that $g(z^*) = g(x^*)$. Consider next the minimization of $g(x)$ over ℓ_i . From the above, the minimum is attained at an intersection point of ℓ_i with some other line ℓ_j , $j = 1, 2, 3$, $j \neq i$. \square

LEMMA 2. *Let S be the interior of a triangle, and let X be a closed connected set in R^2 such that there is no x in R^2 with $\{x\} + X \subseteq S$. Then there is \bar{x} in R^2 such that $\{\bar{x}\} + X$ intersects the three edges of the triangle.*

Proof. Let e_1 , e_2 , and e_3 denote the three edges of the triangle, and let ℓ_1 , ℓ_2 , and ℓ_3 denote the three lines containing the three edges, respectively. Also, for $i = 1, 2, 3$, let ℓ_i^+ be the half plane, determined by ℓ_i , that contains the given triangle. Since X

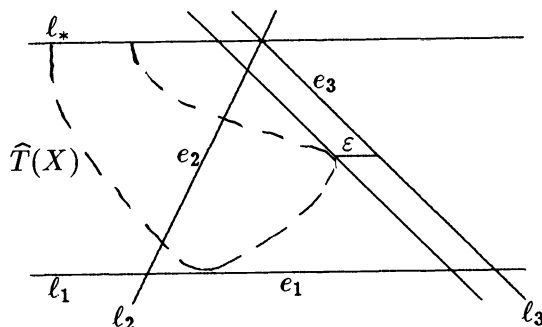


FIG. 2

is closed, there exists a translation of X , say $T(X) = \{z\} + X$ for some $z \in R^2$, that intersects ℓ_1 at some (relative) interior point of e_1 and is contained in ℓ_1^+ .

We first show that $T(X)$ must also intersect either e_2 or e_3 . Indeed, if it does not, then, since $T(X)$ is closed, there is a positive distance between $T(X)$ and $e_2 \cup e_3$. Also, $T(X)$ is included in the triangle since $T(X)$ is connected. Therefore, there is a perturbed translation of $T(X)$ along the normal to ℓ_1 that yields a translation, say $T^1(X) = \{u\} + X$ for some $u \in R^2$, which is included in S .

Thus, suppose that $T(X)$ intersects e_2 but not e_3 . (If it intersects both, the result holds.) Let ℓ_* be a line parallel to ℓ_1 and containing the vertex of the triangle opposing e_1 . Also let ℓ_*^+ denote the half plane, defined by ℓ_* , which contains the triangle.

Without loss of generality, suppose that ℓ_1 is the (horizontal) x_1 -axis in the plane. Define

$$\widehat{T}(X) = T(X) \cap \ell_1^+ \cap \ell_3^+ \cap \ell_*^+.$$

Let $\varepsilon = \text{Minimum}\{y - x_1 \mid (y, x_2) \in e_3, (x_1, x_2) \in \widehat{T}(X)\}$. Due to the fact that $T(X)$ is closed and does not intersect e_3 , ε is well defined and positive. (See Fig. 2.)

Next, we translate $T(X)$ along ℓ_1 toward ℓ_3 by a distance of $\varepsilon > 0$. Let $T^2(X)$ denote the translated set obtained by this move. If $T^2(X)$ intersects e_2 , the result holds. Otherwise, there are two cases to consider. First, suppose that $T^2(X)$ contains a point outside the triangle. Then a contradiction to the connectivity of $T^2(X)$ is easily obtained. Therefore, suppose that $T^2(X)$ is contained in the triangle and does not intersect e_2 .

A perturbed translation of $T^2(X)$ along the bisector of the triangle angle, defined by e_1 and e_3 , moves $T^2(X)$ into S . This completes the proof. \square

THEOREM 4. *Let S be the interior of a triangle. Then any minimum width chord of $\text{Cl}(S)$ is an S -uncontained path of minimum length and minimum measure.*

Proof. It is clear that any minimum width chord of $\text{Cl}(S)$ is an S -uncontained path. Consider an S -uncontained path f . Let $X = f(I)$. From Lemma 2 there exists a translation, say $T(X) = \{z\} + X$ for some $z \in R^2$, that intersects the three edges of $\text{Cl}(S)$, e_1 , e_2 , and e_3 at points x^1 , x^2 , and x^3 , respectively. (The points are not necessarily distinct.) $\lambda_1(X)$, the measure of X (with respect to the one-dimensional Hausdorff measure) is bounded below by the measure of a minimum measure set that (arcwise) connects x^1 , x^2 , and x^3 . It is known [2] that a Euclidean Steiner tree connecting this triplet of points is a minimum measure set. Furthermore, the measure

of such a tree is the sum of the (Euclidean) distances from some point in R^2 to x^1 , x^2 , and x^3 . Using Lemma 1, we note that the measure of the connecting Steiner tree is greater than or equal to the width of $\text{Cl}(S)$. Thus we conclude that the measure of X is not smaller than the width of $\text{Cl}(S)$. Finally, the length of f , $\lambda(f)$ is bounded below by the length of any minimal length path connecting x^1 , x^2 , and x^3 . Thus, $\lambda(f)$ is bounded below by the measure of the above Steiner tree. This completes the proof. \square

THEOREM 5. *Let S be an open bounded convex set in R^2 . Then any minimum width chord of $\text{Cl}(S)$ is an S -uncontained path of minimum length and minimum measure.*

Proof. Let f be an S -uncontained path. To prove that $\lambda(f)$ and $\lambda_1(f(I))$ are both bounded below by the width of $\text{Cl}(S)$, we may suppose that the assumptions of Theorems 2 and 3 are satisfied. Thus we refer to the two cases stated in Theorem 3. If (1) holds, then clearly both $\lambda(f)$ and $\lambda_1(f(I))$ are bounded below by the distance between the two parallel supporting lines. If (2) holds, we apply Theorem 4 to the triangle defined in this case. Again, both $\lambda(f)$ and $\lambda_1(f(I))$ are bounded below by the width of that triangle, which, in turn, is bounded by the width of $\text{Cl}(S)$. This completes the proof. \square

3. Concluding remarks and open problems. We show above that if S is bounded and convex, then an optimal S -uncontained curve is linear and its length is the width of S . Linearity might be lost if S is not convex. In fact, we might have no linear minimal length curve even for the case when S is the union of seven pairwise disjoint open rectangles of the same width. Consider the case when S is a planar polygon, given by an ordered sequence of its vertices. If S is convex, its width can be computed in time that is linear in the number of vertices [15]. When S is not convex, the complexity of determining minimal length or minimal measure curves is still unknown. We suspect that it is NP-hard. We conjecture that minimal curves are piecewise linear and that the number of pieces is polynomial in the number of vertices. If some optimal curve is indeed piecewise linear and a bound on the number of pieces is known a priori, then we can construct a finite scheme to compute minimal curves. The existence of such a scheme follows directly from the theory of Tarski on solvability over real closed fields [16] since the model can be formulated as an algebraic sentence.

We demonstrate above that minimal measure curves are not necessarily simple, i.e., one-to-one. However, we conjecture that there exists a minimal length curve that is simple.

Finally, we mention several extensions and generalizations of the above models. First, we can consider the extension to R^n for $n \geq 3$. We suspect that Theorem 5 holds for this general case as well. Second, we can consider disconnected solution sets. Let X be a closed set in R^2 that contains the origin. Call X an S -uncontained set if for any x in S , the set $\{x\} + X$ intersects the boundary of S . The extended optimization model seeks an S -uncontained compact set of minimal measure with respect to the Hausdorff measure defined above. Since we do not require connectedness of X , we might possibly obtain a solution whose measure is smaller than the solution to Model 2. Indeed, the example in Fig. 3, due to Gal, demonstrates this possibility.

Our model deals with optimal navigation with a compass. We cite in the Introduction several works that discuss navigation models without a compass. To give some mathematical precision to the distinction between a lost swimmer with a compass and a lost swimmer without a compass, consider the following unifying model.

Let \mathcal{R} be a set of transformations of R^2 , and let S be an open set in R^2 . A path

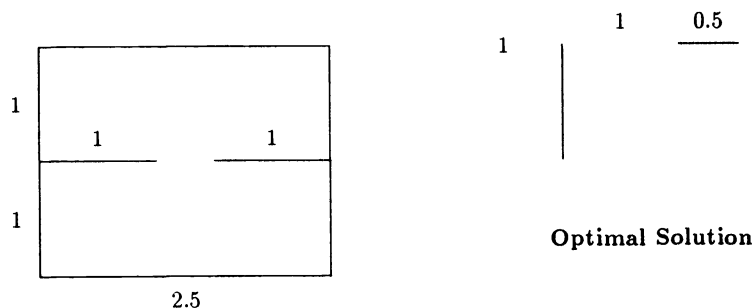


FIG. 3

$f: I \rightarrow R^2$ with $f(0) = 0$ will be called (\mathcal{R}, S) -uncontained if, for every $x \in S$ and a transformation $r \in \mathcal{R}$, the set $\{x\} + r(f(I))$ intersects ∂S .

In our model of the swimmer with the compass, \mathcal{R} consists only of the identity transformation. The problem of the swimmer without a compass is modeled by letting \mathcal{R} be the group of all rotations. Other interesting cases are when \mathcal{R} is the group of all isometries and when $\mathcal{R} = SL_2(R)$, the group of all linear transformations with a determinant being equal to $+1$ or -1 . The existential result of Theorem 1 can easily be generalized to the above examples of \mathcal{R} . Unlike the general result stated in Theorem 5 for convex sets S in our model, finding and verifying an optimal path for a specific set (e.g., a rectangle or a half plane) is fairly involved even while focusing on the swimmer-without-compass model.

We have assumed in all the above models that there is no information about the initial location of the swimmer within the set S . These models must be modified when such information becomes available. For example, if in our original model of navigation with a compass, the swimmer is known to be within a subset S' of S , we require from a path f that only for each $x \in S'$ the set $\{x\} + f(I)$ intersects the boundary of S . The result of Theorem 1 can be extended to this case as well. It is interesting to find sufficient conditions on S and S' that will yield results similar to those stated in Theorem 5.

We have also assumed throughout that the objective is to minimize the maximum path size. In other situations, different objectives may exist, such as minimizing the expected size of the path. In such cases, it is also meaningful to consider probabilistic information on the initial location.

REFERENCES

- [1] R. BELLMAN, *Minimization problem*, Bull. Amer. Math. Soc., 62 (1956), p. 270.
- [2] F. R. CHUNG AND R. L. GRAHAM, *Steiner trees for ladders*, Ann. Discrete Math., 2 (1978), pp. 173–200.
- [3] H. T. CROFT, *Curves intersecting certain sets of great circles on the sphere*, J. London Math. Soc., (2), 1 (1969), pp. 461–469.

- [4] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part I: General Theory*, Wiley-Interscience, New York, 1967.
- [5] H. G. EGGLESTON, *The maximal inradius of the convex cover of a plane connected set of given length*, Proc. London Math. Soc. (3), 45 (1982), pp. 456–478.
- [6] V. FABER, J. MYCIELSKI, AND P. PEDERSEN, *On the shortest curve which meets all the lines which meet a circle*, Ann. Polon. Math., 44 (1984), pp. 249–266.
- [7] V. FABER AND J. MYCIELSKI, *The shortest curve that meets all the lines that meet a convex body*, Amer. Math. Monthly, 93 (1986), pp. 796–801.
- [8] S. GAL, *Search Games*, Academic Press, New York, 1980.
- [9] B. GLUSS, *An alternative solution to the "lost in the sea" problem*, Naval Res. Logist., 8 (1961), pp. 117–121.
- [10] ———, *The minimax path in a search for a circle in a plane*, Naval Res. Logist., 10 (1963), pp. 357–360.
- [11] O. GROSS, *A search problem due to Bellman*, Rand research memorandum RM-1603 Rand Institute, 1955.
- [12] J. R. ISBELL, *An optimal search pattern*, Naval Res. Logist. 4 (1957), pp. 357–359.
- [13] H. JORIS, *Le chasseur perdu dans la forêt: un problème de géométrie plane*, Elem. Math., 35 (1980), pp. 1–14.
- [14] P. A. P. MORAN, *On a problem of S. Ulam*, J. London Math. Soc., 21 (1946), pp. 175–179.
- [15] F. P. PREPARATA AND M. I. SHAMOS, *Computational Geometry—An Introduction*, Springer-Verlag, New York, 1985.
- [16] A. TARSKI, *A Decision Method for Elementary Algebra and Geometry*, 2nd edition, revised, University of California Press, Berkeley, Los Angeles, CA 1951.

FINITE-DIMENSIONAL APPROXIMATIONS OF UNSTABLE INFINITE-DIMENSIONAL SYSTEMS*

G. GU[†], P.P. KHARGONEKAR[‡], E.B. LEE[§] AND P. MISRA[¶]

Abstract. This paper studies approximation of possibly unstable linear time-invariant infinite-dimensional systems. The system transfer function is assumed to be continuous on the imaginary axis with finitely many poles in the open right half plane. A unified approach is proposed for rational approximations of such infinite-dimensional systems. A procedure is developed for constructing a sequence of finite-dimensional approximants, which converges to the given model in the L_∞ norm under a mild frequency domain condition. It is noted that the proposed technique uses only the FFT and singular value decomposition algorithms for obtaining the approximations. Numerical examples are included to illustrate the proposed method.

Key words. finite-dimensional approximations, infinite-dimensional systems, optimal Hankel approximation, balanced realization, discrete Fourier transform

AMS(MOS) subject classifications. 93C25, 93B15, 41A65, 41A20

1. Introduction. Since it is difficult to deal with infinite-dimensional systems directly, often a finite-dimensional approximate model is sought. The problem of approximating infinite-dimensional systems with finite-dimensional ones has been addressed by many authors in both the time domain [2], [8], [14], and the frequency domain [10]–[12], [17], [18], [21], [25], [26]. In this paper, we consider the approximation of possibly *unstable* linear time-invariant infinite-dimensional systems in the *frequency domain*. It is assumed that the transfer function $T(s)$ of the given system is continuous on the imaginary axis, including infinity and has only *finitely* many poles in the open right half plane. The objective is to seek a rational approximant $T_r(s)$, having the same number of unstable poles as $T(s)$, such that $\|T - T_r\|_\infty$ is suitably small. The motivation for this problem comes from feedback design considerations. For example, it follows from Curtain and Glover [5] and Chen and Desoer [4] that a controller stabilizes $T_r(s)$ also stabilizes $T(s)$, provided that $T(s)$ and $T_r(s)$ have the same number of poles in the right half plane and $\|T - T_r\|_\infty$ is suitably small.

One approach for the approximation of such unstable infinite-dimensional systems is to first separate the (finite-dimensional) unstable part by partial fraction expansion, and then consider the approximation of the stable part of system [5], [21]. Although partial fraction expansion is very effective for extracting the unstable part of the given infinite-dimensional system, it requires computation of the right half plane poles of the system. In this paper, we propose an alternative technique for the approximation of unstable infinite-dimensional systems. This work extends to unstable systems

*Received by the editors April 9, 1990; accepted for publication (in revised form) May 14, 1991. This work was supported in part by National Science Foundation grants ECS-9110636, ECS-9001371 and DMS-8722402, by U.S. Air Force Office of Scientific Research contract AFOSR-90-0053, Army Research Office grant DAAL03-90-G-0008, and by WDRC/WPAFB grant F33615-88-C-3605.

[†]Department of Electrical and Computer Engineering, Louisiana State University, Baton Rouge, Louisiana 70809.

[‡]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109.

[§]Department of Electrical Engineering, University of Minnesota, Minneapolis, Minnesota 55455.

[¶]Department of Electrical Engineering, Wright State University, Dayton, Ohio 45435.

certain approximation techniques developed in [11], [17], [18], [25] for stable systems. An important feature of this proposed technique is the unification of approximations of both the stable and the unstable parts of $T(s)$ within a single algorithm. It will be shown that under certain mild conditions, the resulting approximants $T_r(s)$, which have same number of unstable poles as $T(s)$, converge to $T(s)$ in the L_∞ -norm. Moreover, the proposed approximation technique uses only the FFT and singular value decomposition algorithms. Therefore, we expect our method to be preferable from the computational point of view. It should be noted that the fast Fourier transform technique has been used in many problems in the literature on computational complex analysis. See, for example, the survey paper by Henrici [13] and the references therein. Our work shows that these ideas are also very useful in system approximation problems, and lead to concrete convergence results as well as L_∞ -error bounds. Also, the resulting algorithms are computationally very efficient.

It is also noted that other frequency domain approximation techniques such as those developed in [10], [12], [21] might be applicable to the approximation problem considered in this paper. However, we believe that our algorithm is attractive from a computational point of view as compared to some of these algorithms. Also, the extensive work in the Padé approximation literature is potentially applicable to the present problem. However, in this case, convergence and error analysis in the L_∞ norm remains a topic for future research in the context of our problem. Finally, the work of Trefethen [23] is also of interest for our problem. The Caratheodory–Fejer (CF) method proposed in [23] is considered to be very effective for frequency domain approximation [12]. However, it has been recently pointed out by Saff and Totik [22] that the CF method does not always provide a better approximation than partial sums of Fourier series, and there exist functions whose Fourier series converges uniformly but the approximant obtained using the CF method diverges. We would like to emphasize that this should not be taken to imply that the CF method is inferior in comparison to Fourier series. As indicated in [22], the CF method is superior to the partial sum of Fourier series for those functions that are sufficiently smooth.

We believe that the technique proposed in the present paper offers an effective alternative to the techniques that could be derived from the references cited above. Preliminary analysis appears to imply that all these techniques may have different domains of applicability. A comparative study of all these algorithms remains a subject for future work.

The paper is organized as follows. A preliminary result will be presented first for discrete-time systems in §2, which will be used to establish the main result of this paper in §3. Two numerical examples will be given in §4 to illustrate the approximation technique.

2. A preliminary result. Before studying the approximation of unstable infinite-dimensional systems, we will first establish a simple result that will be useful in the next section. Let $G(z)$ be the transfer function of a given linear, time-invariant, finite-dimensional, exponentially stable, discrete time system of McMillan degree n . Suppose that $G(z)$ is given by

$$(2.1) \quad G(z) = \sum_{k=1}^{\infty} g_k z^{-k}, \quad \text{with } g_k \in \mathcal{R}^{p \times m}.$$

Define the partial summation

$$(2.2) \quad S_N(z) := \sum_{k=1}^N g_k z^{-k}.$$

A simple (possibly nonminimal) realization for $S_N(z)$ is given by

$$(2.3) \quad A_N = \begin{bmatrix} 0 & 0 & \dots & \dots & 0 \\ I_m & 0 & 0 & \dots & 0 \\ 0 & I_m & 0 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & I_m & 0 \end{bmatrix}, \quad B_N = \begin{bmatrix} I_m \\ 0 \\ \dots \\ \dots \\ 0 \end{bmatrix}, \quad C_N^T = \begin{bmatrix} g_1^T \\ \dots \\ \dots \\ \dots \\ g_N^T \end{bmatrix}.$$

Since the above realization is controllable, an input normal realization [8], [10] of $S_N(z)$ (which has properties similar to balanced realization) can be easily found by solving two Lyapunov equations. In fact, with the realization as in (2.3), a much simpler algorithm can be used to compute a similarity transformation T (which is a unitary matrix) using only one singular value decomposition (see [11] for more details) such that

$$(2.4) \quad (A_b, B_b, C_b) = (T A_N T^T, T B_N, C_N T^T)$$

is an input normal realization of $S_N(z)$. Now (for $N \geq n$) an approximant $G_n^N(z)$ of degree n can be obtained by direct truncation of the input normal realization (A_b, B_b, C_b) as follows:

$$(2.5) \quad A_n^N = [I_n \quad 0] A_b \begin{bmatrix} I_n \\ 0 \end{bmatrix}, \quad B_n^N = [I_n \quad 0] B_b, \quad C_n^N = C_b \begin{bmatrix} I_n \\ 0 \end{bmatrix}.$$

It is noted that the McMillan degree of $G_n^N(z)$ may be smaller than n . However, if the McMillan degree of S_N is no smaller than n , then the McMillan degree of G_n^N is exactly n . With $G_n^N(z) := C_n^N(zI - A_n^N)^{-1}B_n^N$ as described above, we have the following result.

THEOREM 2.1. *Let $G(z)$ be given as in (2.1). Suppose that $G(z)$ is exponentially stable and has McMillan degree n . Then, $\lim_{N \rightarrow \infty} \|G - G_n^N\|_\infty = 0$, where G_n^N is obtained from (2.2)–(2.5).*

Proof. It is easy to show that (see also [10]), for $k > n$,

$$(2.6) \quad \sigma_k(S_N) \leq \sum_{k=1}^{\infty} \sigma_{\max}(g_{N+k}),$$

where $\sigma_k(S_N)$ is the k th Hankel singular value (in descending order) of $S_N(z)$ and $\sigma_{\max}(g_{N+k})$ is the maximum singular value of $p \times m$ matrix g_{N+k} . Hence,

$$(2.7) \quad \sum_{n+1}^N \sigma_k(S_N) \leq (N - n) \sum_{k=1}^{\infty} \sigma_{\max}(g_{N+k}).$$

Since G_n^N is obtained from (2.2)–(2.5), using the results from Glover [9], Enns [6], and Al-Saggaf and Franklin [1], it is easy to see that

$$(2.8) \quad \|S_N - G_n^N\|_\infty \leq 2 \sum_{k=n+1}^N \sigma_k(S_N).$$

By the triangle inequality and the fact that $\|G - S_N\|_\infty \leq \sum_{k=1}^\infty \sigma_{\max}(g_{N+k})$, it follows from (2.6)–(2.8) that

$$(2.9) \quad \|G - G_n^N\|_\infty \leq \|G - S_N\|_\infty + \|S_N - G_n^N\|_\infty \leq 2(N - n + \tfrac{1}{2}) \sum_{k=1}^\infty \sigma_{\max}(g_{N+k}).$$

By the hypothesis, $G(z)$ admits a minimal realization (A, B, C) where $A \in \mathcal{R}^{n \times n}$, $B \in \mathcal{R}^{n \times m}$ and $C \in \mathcal{R}^{p \times n}$ such that $G(z) = C(zI - A)^{-1}B$ and the spectral radius of A , $\rho(A) < 1$. Therefore,

$$(2.10) \quad \sigma_{\max}(g_k) \leq \alpha \rho(A)^k$$

for some $\alpha > 0$. Hence, the error estimate in (2.9) can then be bounded as

$$(2.11) \quad \|G - G_n^N\|_\infty \leq 2(N - n + \tfrac{1}{2}) \frac{\alpha \rho(A)^N}{1 - \rho(A)}.$$

The condition $\rho(A) < 1$ guarantees that $\lim_{N \rightarrow \infty} \|G - G_n^N\|_\infty = 0$.

Note that the approximant $G_n^N(z)$ can also be obtained from n th-order optimal Hankel approximation of $S_N(z)$ as in (2.2) for which Theorem 2.1 is still true (see Glover [9] and Kung and Lin [16]). However, as N becomes large, the computational burden associated with the Hankel approximation technique would be significantly higher compared to the input-normal-realization-based direct truncation technique. Finally, although the error bound in (2.11) is conservative, it does indicate that the convergence depends directly on the value of $\rho(A)$.

3. Main result. In this section, we consider the approximation of unstable, continuous-time infinite-dimensional system $T(s)$. Let \mathcal{H} and \mathcal{D} denote the open right half plane and the open unit disc, respectively. It is assumed that $T(s) \in L_\infty$ and has only finitely many poles in \mathcal{H} .

As mentioned earlier, one technique for obtaining finite-dimensional approximations is to use the partial fraction expansion to decompose $T(s) = T_s(s) + T_u(s)$ with $T_s(s)$ and $T_u(-s)$ both analytic in \mathcal{H} . Consequently, much of the existing research work concentrates only on the approximation of stable part $T_s(s) = T(s) - T_u(s)$. From the computational point of view, it is preferable to avoid the partial fraction decomposition.

In the rest of this section, we develop a new technique to obtain rational approximations for possibly unstable systems. The transfer function of the given continuous-time infinite-dimensional system is first transformed to a function on the unit circle by means of a bilinear transformation. This transformation preserves the L_∞ norm as well as Hankel singular values [9]. The rational approximant is then obtained by using

the FFT and singular value decomposition algorithms. The resulting approximant is then transformed back to obtain an approximation of the original system by means of the inverse bilinear transformation.

Define the bilinear transformation

$$(3.1) \quad s := \lambda \frac{1-z}{1+z} \quad \text{or} \quad z := \frac{\lambda-s}{\lambda+s},$$

which is a conformal mapping from \mathcal{H} to \mathcal{D} . Next, define

$$(3.2) \quad F(z) := T \left(\lambda \frac{1-z}{1+z} \right).$$

Then, $F(z) = \sum_{k=-\infty}^{\infty} f_k z^k = F_s(z) + F_u(z)$, (which converges in the L_2 -sense) with

$$(3.3) \quad F_u(z) = \sum_{k=1}^{\infty} f_{-k} z^{-k} \quad \text{and} \quad F_s(z) = \sum_{k=0}^{\infty} f_k z^k.$$

Clearly, $F_u(z)$ and $T_u(s)$ have the same McMillan degree which by assumption is finite. Furthermore, since the bilinear transformation does not change the Hankel singular values of the original transfer function as shown in [9], the Hankel singular values of $F_u(z)$ are exactly the same as those of $T_u(s)$. Therefore, if the sequence $\{f_{-k}\}_{k=1}^{\infty}$ is known precisely, $F_u(z)$ can be reconstructed using a number of different techniques from the realization theory literature. A problem arises, since we would like to avoid computing the sequence $\{f_k\}$ exactly. Let us define a $2M$ -point inverse discrete Fourier transform as follows to compute $\{f_k\}$ approximately:

$$(3.4) \quad f_M(k) = \frac{1}{2M} \sum_{r=-M}^{M-1} F(W_{2M}^r) W_{2M}^{-rk}, \quad k = -M, -M+1, \dots, M-1,$$

where $W_{2M} = e^{j\pi/M}$. The sequence $\{f_M(k)\}$ can then be used as an approximation for $\{f_k\}$.

The DFT-based approximation has been studied in [11] for stable infinite-dimensional systems, and the convergence, as well as the error bounds, are established for a class of infinite-dimensional systems. Here, we concentrate on the approximation of the unstable part of the system and obtain some similar results. We first state a lemma based on which the main result of the paper will be obtained.

LEMMA 3.1. *Let $F(z)$ be defined as in (3.2) and let $F_u(z)$ be of finite McMillan degree. Suppose that $dF(e^{j\omega})/de^{j\omega} \in L_2[0, 2\pi]$. Then,*

$$(i) \quad \{\|f_k\|\} \in \ell^1, \text{ (that is, } \sum_{k=-\infty}^{\infty} \|f_k\| < \infty) \text{ and}$$

$$(ii) \quad f_M(k) = \sum_{L=-\infty}^{\infty} f_{2LM+k},$$

where f_k and $f_M(k)$ are defined by (3.3) and (3.4), respectively.

This result is quite well known. See, for example, [15], [13].

Note that since $F_u(z)$ is analytic on unit circle and its McMillan degree is finite, the condition $dF(e^{j\omega})/de^{j\omega} \in L_2[0, 2\pi]$ is, in fact, equivalent to $dF_s(e^{j\omega})/de^{j\omega} \in$

$L_2[0, 2\pi]$. Hence, $F(z)$ (or $F_s(z)$) is continuous on the unit circle. By (3.2), the continuity of $F(z)$ on unit circle is equivalent to the continuity of $T(s)$ on the extended $j\omega$ axis. Therefore the hypothesis in Lemma 3.1 implies that the transfer function $T(s)$ admits rational approximants that converge in L_∞ -norm to T . Our results in this paper, in fact, give a constructive procedure for obtaining such approximants.

THEOREM 3.2. *Let $F(z)$ be defined as in (3.2) and $F_u(z)$ have McMillan degree n . Define $S_N^M(z) := \sum_{k=1}^N f_M(-k)z^{-k}$, with $f_M(k)$ as in (3.4), $N > n$. Suppose that $dF(e^{j\omega})/de^{j\omega} \in L_2[0, 2\pi]$. Then*

$$(3.5) \quad \lim_{\sqrt{M} \geq N \rightarrow \infty} \|F_u - F_{u;n}^{M;N}\|_\infty = 0,$$

where $F_{u;n}^{M;N}(z)$ is an n th-order approximant of $S_N^M(z)$ obtained using the balanced truncation scheme described in §2 (or the optimal Hankel approximation method).

Proof. By the triangle inequality,

$$(3.6) \quad \|F_u - F_{u;n}^{M;N}\|_\infty \leq \|F_u - S_N\|_\infty + \|S_N - S_N^M\|_\infty + \|S_N^M - F_{u;n}^{M;N}\|_\infty,$$

where $S_N(z) = \sum_{k=1}^N f_{-k}z^{-k}$. We will show that the three terms on the right-hand side of (3.6) approach zero as $\sqrt{M} \geq N \rightarrow \infty$.

Indeed, because $F_u(z)$ is rational and has all its poles in \mathcal{D} , it is easy to see that

$$(3.7) \quad \lim_{N \rightarrow \infty} \|F_u - S_N\|_\infty = 0.$$

Furthermore, since $\|z^{-k}\|_\infty = 1$, we have that $\|S_N - S_N^M\|_\infty \leq \sum_{k=1}^N \sigma_{\max}(f_M(-k) - f_{-k})$. Lemma 3.1 implies that

$$(3.8) \quad f_\Delta(k) = f_M(k) - f_k = \sum_{L \neq 0} f_{2LM+k},$$

where the summation is with respect to L and k is fixed. Hence,

$$(3.9) \quad \|S_N - S_N^M\|_\infty \leq \sum_{k=1}^N \sigma_{\max}(f_\Delta(k)) \leq \sum_{k=1}^\infty \{\sigma_{\max}(f_{M+k-1}) + \sigma_{\max}(f_{-M-k})\} \rightarrow 0$$

as $M \rightarrow \infty$. Finally, the third term on the right-hand side of (3.6) is bounded by

$$(3.10) \quad \|S_N^M - F_{u;n}^{M;N}\|_\infty \leq \beta \sum_{i=n+1}^N \sigma_i(S_N^M),$$

where $\beta = 1$ if $F_{u;n}^{M;N}$ is obtained from optimal Hankel approximation of S_N^M [9] and $\beta = 2$, if $F_{u;n}^{M;N}$ is obtained from the reconstruction scheme in §2 or the balanced realization technique for S_N^M [1]. Now

$$(3.11) \quad S_N^M(z) = \sum_{k=1}^N f_M(-k)z^{-k} = S_N(z) + \sum_{k=1}^N f_\Delta(-k)z^{-k}.$$

By singular value perturbation results and Theorem 2.1, we get

$$\begin{aligned}
 \sigma_i(S_N^M) &\leq \sigma_i(S_N) + \sum_{k=1}^N \sigma_{\max}(f_{\Delta}(-k)) \\
 &\leq \sum_{k=1}^{\infty} \sigma_{\max}(f_{-N-k}) + \sum_{k=1}^N \sum_{L \neq 0} \sigma_{\max}(f_{2LM-k}),
 \end{aligned}
 \tag{3.12}$$

whenever $i > n$. Note that in deriving the above inequality, (2.6) is used with g_k replaced by f_{-k} . Taking sum of σ_i for $n+1 \leq i \leq N$ above, we get

$$\|S_N^M - F_{u;n}^{M;N}\|_{\infty} \leq \beta(N-n) \sum_{k=1}^{\infty} \{2\sigma_{\max}(f_{-N-k}) + \sigma_{\max}(f_{M+k})\}.
 \tag{3.13}$$

Since the derivative of $F(z)$ is absolutely square-integrable on the unit circle and its unstable part is finite-dimensional, the derivative of $F_s(z)$ is also absolutely square-integrable on the unit circle, which implies that

$$f_{M+k} = \frac{\hat{f}_{M+k}}{M+k} \quad \text{and} \quad \sum_{k=1}^{\infty} \sigma_{\max}(\hat{f}_k)^2 < \infty,
 \tag{3.14}$$

where \hat{f}_k is the k th Fourier series coefficient of $dF(z)/dz$. Moreover, since the unstable part has only n poles on open unit disc, $\sigma_{\max}(f_{-k}) \leq \alpha_o \rho_o^k$ for some $\alpha_o > 0$ and $\rho_o < 1$, where $k > 0$. Therefore, using the Schwarz inequality, we have

$$\|S_N^M - F_{u;n}^{M;N}\|_{\infty} \leq 2\beta(N-n) \frac{\alpha_o \rho_o^N}{1-\rho_o} + \frac{\beta(N-n)}{\sqrt{M}} \sqrt{\sum_{k=1}^{\infty} \sigma_{\max}(\hat{f}_{M+k})^2} \rightarrow 0,
 \tag{3.15}$$

as $\sqrt{M} \geq N \rightarrow \infty$. The proof is now complete using (3.6)–(3.9) and (3.15).

Remark 3.1. It is important to note that n , the McMillan degree of $F_u(z)$, may not be known in advance. However, from Theorem 3.2 (also (3.12)), the first n Hankel singular values of $S_N^M(z)$ converge to the true Hankel singular values of $F_u(z)$ and the rest of the Hankel singular values converge to zero as $(M, N) \rightarrow (\infty, \infty)$, with $M > N$. Therefore, as M, N are both large, a gap between $\sigma_n(S_N^M)$ and $\sigma_{n+1}(S_N^M)$ would be significant if $\sigma_n(F_u)$ is not too small. In this case, the McMillan degree of $F_u(z)$ can also be identified in the approximation process.

Since $F(z)$ is given by (3.2), the frequency domain condition $dF(e^{j\omega})/de^{j\omega} \in L_2[0, 2\pi]$ is, in fact, equivalent to $(\lambda - j\omega)(dT(j\omega)/dj\omega) \in L_2[-\infty, \infty]$ (see [24]). This condition is difficult to verify in general. However, for a class of time delay systems, we can state the following.

PROPOSITION 3.3. *Let $T(s)$ be a transfer function of the form*

$$T(s) = \frac{\sum_{k=0}^m Q_k(s) \exp(-h_k s)}{s^n + \sum_{k=0}^n p_k(s) \exp(-\tau_k s)},
 \tag{3.16}$$

where $p_k(s)$ is a scalar polynomial of s , $Q_k(s)$ is a polynomial matrix of size $m \times r$, and $0 \leq h_0 \leq h_1 \leq \dots \leq h_m, \tau_k > 0$ for $0 \leq k \leq n$. Let $d_k = \deg(Q_k(s))$ and

$\delta_k = \deg(p_k(s))$. It is assumed that $d_k < n$ and $\delta_k < n$. Then $(\lambda - j\omega)(dT(j\omega)/dj\omega) \in L_2[-\infty, \infty]$, if the following statements hold: (1) $T(s)$ is continuous on imaginary axis; and (2) (i) $d_k \leq n - 1$, if $h_k = 0$; (ii) $d_k < n - 1$, if $h_k \neq 0$.

We omit the proof of the above proposition, as it is an easy extension of a result in [11]. To conclude our results, we summarize the following algorithm for rational approximation of unstable part of the given infinite-dimensional system.

Algorithm 3.1 (Rational approximation).

Step 1: For a given unstable infinite-dimensional transfer function $T(s)$, verify first if $(\lambda - j\omega)(dT(j\omega)/dj\omega) \in L_2[-\infty, \infty]$ and choose $\lambda > 0$ to find $F(z)$ as in (3.2);

Step 2: Use $2M$ -point inverse FFT algorithm to compute $f_M(k)$ as defined in (3.4);

Step 3: Compute Hankel singular values of S_N^M as defined in (3.11) with $N^2 \leq M$, N large enough, and estimate n : the number of unstable poles of $T(s)$;

Step 4: Apply (2.3)–(2.5) to $S_N^M(z)$ to obtain $F_{u;n}^{M;N}(z) = C_n^N(zI - A_n^N)^{-1}B_n^N$;

Step 5: Use bilinear transform (3.1) to obtain an approximation for the unstable part of $T(s)$.

End

As discussed earlier, the hypothesis in Theorem 3.2 (which is same as in Lemma 3.1) implies that $dF_s(e^{j\omega})/de^{j\omega} \in L_2[0, 2\pi]$. The approximation of such $F_s(z)$ using $\{f_M(k)\}_{k=0}^\infty$ has been studied previously in [11], where some convergence results and the error bounds have been established. Hence, we will only briefly describe the approximation of the stable part of $F(z)$ below.

Define the partial summation as the approximant of the stable part

$$(3.17) \quad S^{M;L}(z) = \sum_{k=0}^L f_M(k)z^k = J_L + H_L(z^{-1}I - F_L)^{-1}G_L,$$

where the realization (F_L, G_L, H_L, J_L) is similar to (2.3). Therefore, the above realization can be easily converted into an input normal realization by computing only one singular value decomposition. The rational approximant of the stable part of McMillan degree no larger than ℓ can then be obtained by direct truncation as in (2.5), which is denoted as

$$(3.18) \quad F_{s;\ell}^{M;L}(z) = J_\ell^L + H_\ell^L(z^{-1}I - F_\ell^L)^{-1}G_\ell^L.$$

It has been established in [11] that if the conditions in Theorem 3.2 are true, then

$$(3.19) \quad \lim_{M \geq L \geq \ell \rightarrow \infty} F_{s;\ell}^{M;L}(z) = F_s(z), \quad \forall z \in \mathcal{D}.$$

The procedure described above is similar to the approximation of unstable part except that N is replaced by L , and n is replaced by ℓ , and while n is kept fixed, $\ell \rightarrow \infty$. Therefore, the approximation of both stable and unstable part of $T(s)$ can be handled with Algorithm 3.1. The final approximant of $F(z)$ can then be obtained as

$$(3.20) \quad F_r(z) = F_{s;\ell}^{M;L}(z) + F_{u;n}^{M;N}(z).$$

Finally, the finite-dimensional approximation given by $T_r(s) := F_r((\lambda - s)/(\lambda + s))$ is an approximation of $T(s)$. With these definitions, we have the main result of the paper.

THEOREM 3.4. *Let $T(s)$ be the transfer function of a given infinite-dimensional system having finitely many poles on open right half plane. Assume that $T(s)$ is continuous on extended $j\omega$ -axis and $(\lambda - j\omega)dT(j\omega)/dj\omega \in L_2(-\infty, \infty)$. Then, with $T_r(s)$ obtained from the above approximation procedure, $\lim \|T - T_r\|_\infty = 0$, as $M \geq L \geq \ell \rightarrow \infty$ and $\sqrt{M} \geq N \rightarrow \infty$.*

It is noted that the bilinear transform does not change the L_∞ -norm of the transfer function and thus $\|T - T_r\|_\infty = \|F - F_r\|_\infty$.

Remark 3.2. We would like to indicate further that as the approximate model is used for feedback control system design, the following condition should be satisfied to ensure the existence of stabilizing feedback compensator [5]:

$$(3.21) \quad \|T - T_r\|_\infty < \sigma_n(F_{u;n}^{M;N}) = \sigma_{\min}(F_{u;n}^{M;N}).$$

This is due to the fact that the bilinear transform does not change the Hankel singular values.

4. Illustrative examples. To illustrate the approximation technique proposed in this paper, two examples are presented below.

Example 4.1. Consider the following transfer function:

$$(4.1) \quad T(s) = \frac{20(6e^{-2s} + 2e^{-s} - 6)}{6s^2 + (6e^{-2s} - 2e^{-s} - 66)s - (2e^{-3s} + 30e^{-2s} - 12e^{-s} - 180)}.$$

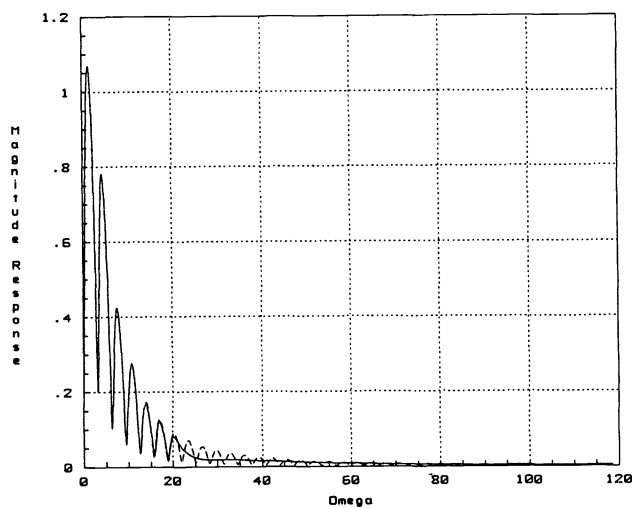
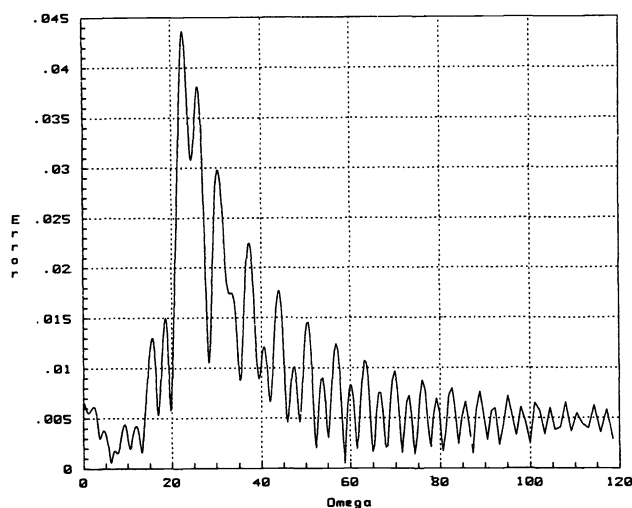
It is easy to verify that the above $T(s)$ is continuous on imaginary axis. Furthermore, the convergence condition in Theorem 3.2 is true in light of Proposition 3.3. If the partial fraction is used for approximation, the poles of $T(s)$ as well as the residues must be computed, which, computationally, is very demanding. Using Algorithm 3.1, we applied bilinear transform (3.1) with $\lambda = 10$. A 2048-point FFT program is used to compute the sequence $\{f_M(k)\}$ as defined in (3.4). The partial sum $S_N^M(z)$ is obtained with $N = 9$. The gap between the second and third Hankel singular values was significant, suggesting that the number of unstable poles of $T(s)$ is $n = 2$. The approximate unstable part of the system is finally obtained as

$$(4.2) \quad \hat{T}_u(s) = \frac{-0.0448(s + 439.34)}{(s - 5.0035)(s - 5.9981)}.$$

It is noted that the exact unstable poles are 5.002224 and 5.999994, which are very close to the poles of $\hat{T}_u(s)$ above.

We would like to mention that this particular transfer function has a very rich frequency response (see the dash line curve in Fig. 4.1). Hence, it is not easy to find a simple finite-dimensional approximation for this transfer function. We have also used Algorithm 3.1 (with necessary modification) for approximation of the stable part with $L = 45$ and $\ell = 15$. The approximant $\hat{T}_s(s)$ of degree 15 is obtained. By setting $T_r(s) = \hat{T}_s(s) + \hat{T}_u(s)$ as the approximant, a satisfactory result is achieved. The magnitude frequency response of the approximant $T_r(s)$ is plotted in solid line in Fig. 4.1. It is seen that the frequency response of $T_r(s)$ matches very well with that of $T(s)$ for the first seven peaks. The frequency response of the error function can be found in Fig. 4.2. Note that

$$(4.3) \quad \|T - T_r\|_\infty = 0.0437 < \sigma_{\min}(\hat{T}_u) = 0.0689.$$

FIG. 4.1. Frequency response of $\hat{T}(s)$ and $T(s)$.FIG. 4.2. Frequency response of $\hat{T}(s) - T(s)$.

Therefore, the existence of feedback compensators, which stabilizes both $T_r(s)$ and $T(s)$, is guaranteed. The approximant $T_r(s)$ in fractional form is listed in Table 4.1.

Example 4.2. Consider the system described by delay-differential equation

$$\begin{aligned}
 \dot{x}(t) &= A_1 x(t) + A_2 x(t-1) + Bu(t) \\
 (4.4) \quad &= \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ -1 & 0 & -1 & 1 \\ 0 & 1 & -1 & 0 \end{bmatrix} x(t) + \begin{bmatrix} -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} x(t-1) + \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} u(t).
 \end{aligned}$$

The above system was used in Fiagbedzi and Pearson [7], where stabilization

TABLE 4.1
Coefficients of $\hat{T}_s(s)$ in Example 4.1

Power of s	Numerator coefficients	Denominator coefficients
17	-4.674443306368986e-03	1.000000000000000e+00
16	4.239380836236473e-01	2.344791598941533e+01
15	-4.521836609299845e+01	1.291295941920766e+03
14	3.466157426093782e+02	2.081800085936672e+04
13	-6.582229625369409e+04	5.962725161939924e+05
12	9.097441642147570e+04	6.002018359817575e+06
11	-3.713504005871573e+07	1.216507871727509e+08
10	6.293813076095961e+06	5.190579448022859e+08
9	-9.927606421876093e+09	1.075305465480490e+10
8	-6.799049187753367e+08	-3.125834632061094e+10
7	-1.313153757820815e+12	3.399723344157356e+11
6	-1.029096097997340e+11	-5.810536260868385e+12
5	-8.195766310713688e+13	6.376021158219976e+12
4	-3.100690880350008e+12	-1.624831770313712e+14
3	-2.059046683673155e+15	5.092921660281645e+14
2	-5.551124575148588e+13	3.600520359638086e+14
1	-1.354380608580234e+16	8.398888911991564e+15
0	2.229241917067798e+15	9.157772117540838e+15

with state feedback was investigated. The synthesis technique proposed in [7] involves the computation of undesirable modes of the system. The state feedback law is then designed to shift the undesirable modes to the left of $\text{Re}(s) = \nu_o$, where $\nu_o < 0$ represents the stability margin of the closed-loop system. For the above system, $\nu_o = -1$ was chosen in [7]. We demonstrate that the proposed approximation technique can also be used to compute the poles to the right side of $\text{Re}(s) = \nu_o$.

First, we choose $C = B^T$, so that

$$(4.5) \quad \begin{aligned} T(s) &= C(sI - A)^{-1}B \\ &= \frac{s^3 + (1 + e^{-s})s^2 + (1 + 2e^{-s})s + e^{-s}}{s^4 + (1 + e^{-s})s^3 + 2(1 + e^{-s})s^2 + (1 + 2e^{-s})s + 2e^{-s}}, \end{aligned}$$

where $A = A_1 + A_2e^{-s}$. It is not difficult to show that with $C = B^T$, the system is both controllable and observable. Furthermore, using Proposition 3.3, the convergence conditions as in Theorem 3.2 are satisfied. Next, we take $\tilde{s} = s + 1$ and determine the unstable part of $T_a(\tilde{s}) = C(\tilde{s}I - A - I)^{-1}B$. Using Algorithm 3.1, we select $\lambda = 1$ for bilinear transform and use a 2048-point inverse FFT to compute $\{f_M(k)\}$. Since the fifth Hankel singular value of $S_{15}^{1024}(z)$ is very small, it is clear that the number of unstable poles of $T_a(\tilde{s})$ is $n = 4$. With the model reduction scheme described in §2, we obtain a fourth-order approximant $F_{u;n}^{M;N}(z)$. Based on the rational approximant of unstable part, we finally computed the poles of $T(s)$ on the right of $\text{Re}(s) = -1$ approximately, as below:

$$\{-0.186364675 \pm j0.91770066797; 0.11438695855 \pm j1.517680152\}.$$

The above poles are very close to the ones in [7], which are computed using the algorithm in Manitius et al. [19] (within 11-digit of exact poles).

Remark 4.1. It should be emphasized that in the above computation, we used only an inverse FFT and a singular value decomposition program, whereas the algorithm in [19] involves searching for poles on each rectangular region of the s -plane

by computing Cauchy index with contour integral and then applying the numerical procedure to find the roots of the exponential polynomial in that particular region. The proposed method, therefore, provides significant computational saving.

Remark 4.2. The selection of parameter λ in bilinear transformation in (3.1) is important in computing the approximant. Extensive experimental experience shows that selecting λ as the bandwidth of $T(s)$ often yields better numerical results.

5. Concluding remarks. In this paper, we proposed a systematic procedure for rational approximation of unstable infinite-dimensional systems with finitely many right half plane poles (McMillan degree n). Convergence results for rational approximation from truncated Fourier series expansion of the given system transfer function were established. A computational procedure using FFT and singular value decomposition algorithms was outlined. The proposed technique is numerically more reliable and computationally more efficient than the existing procedures to achieve the same objective. Numerical examples illustrated the performance of the proposed technique.

REFERENCES

- [1] U. M. AL-SAGGAF AND G. F. FRANKLIN, *An error bound for a discrete reduced order model of a linear multivariable systems*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 815–819.
- [2] H. T. BANKS AND F. KAPPEL, *Spline approximations for functional differential equations*, J. Differential Equations, 34 (1979), pp. 496–522.
- [3] A. BULTHEEL, *Laurent Series and Padé Approximation*, Birkhauser, Basel, Boston, 1987.
- [4] M. J. CHEN AND C. A. DESOER, *Necessary and sufficient conditions for robust stability of linear distributed systems*, Internat. J. Control, 35 (1982), pp. 255–267.
- [5] R. F. CURTAIN AND K. GLOVER, *Robust stabilization of infinite-dimensional systems by finite-dimensional controllers*, Systems and Control Lett., 7 (1986), pp. 41–47.
- [6] D. ENNS, *Model reduction for control system design*, Ph.D. dissertation, Department of Aeronautics and Astronautics, Stanford University, Stanford, CA, 1984.
- [7] Y. A. FIAGBEDZI AND A. E. PEARSON, *Feedback stabilization of linear autonomous time lag systems*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 847–854.
- [8] J. S. GIBSON, *Linear-quadratic optimal control of hereditary differential systems: infinite dimensional Riccati equations and numerical approximation*, SIAM. J. Control Optim., 21 (1983), pp. 95–139.
- [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] K. GLOVER, R. F. CURTAIN, AND J. R. PARTINGTON, *Realization and approximation of linear infinite dimensional systems with error bounds*, SIAM J. Control Optim., 26 (1988), pp. 863–898.
- [11] G. GU, P. P. KHARGONEKAR AND E. B. LEE, *Approximation of infinite dimensional systems*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 610–618.
- [12] J. W. HELTON AND A. SIDERIS, *Frequency response algorithm for H^∞ optimization with time domain constraints*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 427–434.
- [13] P. HENRICI, *Fast Fourier methods in computational complex analysis*, SIAM Rev., 21 (1979), pp. 481–527.
- [14] K. ITÔ AND R. TEGALS, *Legendre–Tau approximation for functional differential equations*, SIAM J. Control Optim., 24 (1986), pp. 737–759.
- [15] Y. KATZNELSON, *An Introduction to Harmonic Analysis*, Dover, New York, 1976.
- [16] S.-Y. KUNG AND D. W. LIN, *Optimal Hankel-norm model reductions: multivariable systems*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 832–852.
- [17] P. M. MAKILA, *Laguerre series approximation of infinite-dimensional systems*, Internal Report, Dept. of Chemical Engineering, Swedish University of Abo, Finland, 1988.
- [18] ———, *Approximation of stable systems by Laguerre filters*, Automatica, 26 (1990), pp. 333–345.
- [19] A. MANITIUS, H. TRAN, G. PAYRE, AND R. ROY, *Computation of eigenvalues associated with functional differential equations*, SIAM J. Sci. Stat. Comp., 8 (1987), pp. 222–247.

- [20] B. C. MOORE, *Principal component analysis in linear systems: controllability, observability and model reduction*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 17–32.
- [21] J. R. PARTINGTON, K. GLOVER, H. J. ZWART, AND R. F. CURTAIN, *L_∞ -approximation and nuclearity of delay systems*, Systems Control Lett., 10 (1988), pp. 59–65.
- [22] E. B. SAFF AND V. TOTIK, *Limitations of Caratheodory–Fejer’s method for polynomial approximation*, J. Approx. Theory, 58 (1989), pp. 284–296.
- [23] L. N. TREFETHEN, *Rational Chebyshev approximation on the unit disk*, Numer. Math., 37 (1981), pp. 297–320.
- [24] M. VIDYASAGAR, *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA., 1985.
- [25] N. E. WU, *A factorization approach to control synthesis of distributed linear systems*, Ph. D. Dissertation, Center for Control Science and Dynamical Systems, University of Minnesota, Minneapolis, MN, 1987.
- [26] N. E. WU AND G. GU, *Discrete Fourier transform and H^∞ approximation*, IEEE Trans. Automat. Control, 35 (1990), pp. 1044–1046.

SOME REMARKS ON THE RICCATI EQUATION ARISING IN AN OPTIMAL CONTROL PROBLEM WITH STATE- AND CONTROL-DEPENDENT NOISE*

GIANMARIO TESSITORE†

Abstract. This paper solves a quadratic optimal control for a linear stochastic evolution equation with unbounded coefficients. It is assumed that the stochastic noise depends both on the state and on the control. The dynamic programming approach is used and attention is focused on the Riccati equation. In §§5 and 6 some attractivity and maximality properties of the solutions of the algebraic Riccati equation are proved and it is shown that, in some special cases, there exists a maximal solution.

Key words. analytic semigroups, regularly dissipative operators, Itô stochastic calculus, dynamic programming, Riccati equation

AMS(MOS) subject classifications. 49A22, 49A60, 49C20, 60H15

Introduction. In this work we consider a quadratic optimal control problem for a system governed by the following “state equation”:

$$dy = (Ay + Bu)dt + CydW_t^{(1)} + DudW_t^{(2)},$$

where y takes values in a Hilbert space H and $W^{(1)}, W^{(2)}$ are independent Wiener processes taking values in the Hilbert spaces K^1 and K^2 , respectively. We assume that A is a regularly dissipative operator (see [15]) in H and allow the linear operator C to be unbounded in H .

We solve our optimal control problem by the dynamic programming approach; this technique has been used in several similar situations (see [4]–[6], [7], [9], [10], [14], [16]).

Special cases of the above problem have been treated by several authors. For instance, the finite-dimensional case is examined in [8], while in [9] and [10] the problem is treated when H is an infinite-dimensional Hilbert space and C is a bounded linear operator. The existence and uniqueness of the solution of a large class of linear stochastic differential equations that includes our “state equation” when $\mathcal{D}(A) = \mathcal{D}(A^*)$, is proved in [3]. In [5] this result is used to treat the above-mentioned control problem when $D = 0$ and A is self-adjoint.

In this paper we work under general hypotheses: A is only assumed to be regularly dissipative and, as is usually done in the finite-dimensional case (see [12]), the presence of a control-dependent stochastic noise is allowed. However, only distributed controls are treated here; for some results concerning the boundary control case, see [7].

The model we use here covers a wide class of parabolic stochastic partial differential equations (p.d.e.’s) with strongly elliptic differential operators A . For example, the equation

$$dy = \left[\sum_{i,j=1}^d \frac{\partial}{\partial x_i} \left(a_{i,j}(x) \frac{\partial y}{\partial x_j} \right) + a(x)y + b(x)u(t, x) \right] dt$$

*Received by the editors June 11, 1990; accepted for publication (in revised form) May 8, 1991.

†Scuola Normale Superiore, Piazza dei Cavalieri no. 7, 56126 Pisa, Italy.

$$+ \left[\sum_{i=1}^d c_i(x) \frac{\partial y}{\partial x_i} + c(x)y \right] dw_t^{(1)} + u(t, x) dw_t^{(2)}, \quad x \in \mathcal{O};$$

$$\frac{\partial y}{\partial \nu_a}(t, x) = 0, \quad x \in \partial \mathcal{O};$$

$$y(0, x) = y_0(x), \quad x \in \mathcal{O}$$

(where \mathcal{O} is a bounded domain in \mathbb{R}^d with smooth boundary) can be, under usual hypotheses of ellipticity and regularity on the coefficients, reduced to our abstract model (see §7); on the contrary, the assumptions required in [3] and [5] are not fulfilled.

The presence of the unbounded linear operator C causes some technical difficulties. In fact, we must apply the Itô formula, which requires the use of classical solutions both for the state and for the Riccati equations. For this reason, we are naturally led to introduce approximating problems in which the unbounded operators are replaced by suitable Yoshida approximations. Then, we must show that the solutions of the state and Riccati equations corresponding to these approximating problems converge to the solutions of the original state and Riccati equations. Moreover, the “state equation” is solved by a fixed-point argument in a space of stochastic processes taking values in a Hilbert space $V \neq H$, whereas the approximating equations are solved in a space of stochastic processes taking values in H . This discrepancy in the function spaces yields some technical complications in proving the above-mentioned convergence, which are here overcome by splitting the approximation in two steps (see §1).

The Riccati equation corresponding to the optimal control problem considered in this work is the following (see [9] and [10]):

$$(\mathcal{R.E.}) \quad P' = A^*P + PA + \sum_{i=1}^{+\infty} \lambda_i C_i^* P C_i - PB \left[I + \sum_{i=1}^{+\infty} \nu_i D_i^* P D_i \right]^{-1} B^*P + S,$$

$$P(0) = P_0.$$

When no control-dependent noise is considered, the term

$$\gamma(P) = PB \left[I + \sum_{i=1}^{+\infty} \nu_i D_i^* P D_i \right]^{-1} B^*P$$

is replaced by the much simpler term PBB^*P (see [4] and [4]), and the right-hand side of $(\mathcal{R.E.})$ reduces to a locally Lipschitzian map on the real Banach space of all bounded and self-adjoint operators in H (this space will be denoted by $\Sigma(H)$).

Hence $(\mathcal{R.E.})$ can be, in this special case, locally solved by a fixed-point argument in the space of all strongly continuous functions taking values into $\Sigma(H)$ (this space will be denoted by $\mathcal{C}_u(0, T, \Sigma(H))$); to be exact, we must say that this is true only if we consider “mild” solutions. However, in the general case, γ is not defined in all $\Sigma(H)$. In [10] the problem is avoided using the monotone sequence given by the solutions of a suitable class of differential equations. This method, however, seems hard to generalize due to the unboundedness of the operators C_i ’s. In this paper we use a fixed-point argument in a suitable closed subset of $\mathcal{C}_u(0, T, \Sigma(H))$ (see §2); in this way we can immediately prove not only that a local solution of $(\mathcal{R.E.})$ does exist (and is unique) but also that it is the limit of the solutions of the Riccati equations corresponding to the above-mentioned approximating problems.

Once the state and the Riccati equations have been solved, it is possible to show the existence of an optimal control, both for the finite and for infinite horizon problem, as in [4] and [5]. The infinite horizon case is here treated under a stabilizability hypothesis that is weaker than the one used in [5] and [10] (for an extensive discussion of stabilizability problems in the deterministic case, see [14]).

In §5 we show some attractivity and maximality properties of a class of solutions of the corresponding algebraic Riccati equation

$$(\mathcal{A.R.E.}) \quad A^*X + XA + \sum_{i=1}^{+\infty} \lambda_i C_i^* X C_i - XB \left[I + \sum_{i=1}^{+\infty} \nu_i D_i^* X D_i \right]^{-1} B^*X + S = 0$$

(see [17] for similar results in the deterministic case and [5] for some earlier stochastic generalizations). Using these properties, we show (see §6), under suitable stabilizability hypotheses, the existence of a maximal solution of $(\mathcal{A.R.E.})$, and we show that it is related to a special kind of control problem with a restricted class of admissible controls (for the deterministic case, see [2]).

It seems that our approach could work in more general situations including boundary control problems; the extension of the previous results will be treated in a future work.

1. State equation. The control problem that we deal with in this work is given by the following state equation:

$$(\mathcal{S.E.}) \quad \begin{aligned} dy &= (Ay + Bu)dt + CydW_t^{(1)} + DudyW_t^{(2)}, \\ y(0) &= x \end{aligned}$$

and by one of the following cost functionals:

$$\begin{aligned} J_T(x, u) &= \mathbf{E} \int_0^T \left(\|\sqrt{S}y(s)\|^2 + \|u(s)\|^2 \right) ds + \mathbf{E} \langle P_0 y(T), y(T) \rangle \quad (\text{finite horizon case}), \\ J_\infty(x, u) &= \mathbf{E} \int_0^{+\infty} \left(\|\sqrt{S}y(s)\|^2 + \|u(s)\|^2 \right) ds \quad (\text{infinite horizon case}), \end{aligned}$$

where, in both cases, y is the solution of $(\mathcal{S.E.})$ (in a mild sense made precise below).

We suppose that A is a regularly dissipative operator. This means (see [15]) that there exist two Hilbert spaces, V (norm $|\cdot|$) and H (norm $\|\cdot\|$) with V continuously and densely imbedded in H and bilinear form $a(\cdot, \cdot)$ defined and continuous in V , verifying $-a(v, v) \geq c|v|^2 - \bar{\omega}\|v\|^2$ for some $c > 0$, $\bar{\omega} \geq 0$, and all $v \in V$, and A is defined as follows:

$$\begin{aligned} \mathcal{D}(A) &= \{x \in V \text{ s.t. the map } y \rightarrow a(x, y) \text{ is continuous in } H\}, \\ \forall x \in \mathcal{D}(A) \quad Ax &\text{ is the only element in } H \text{ s.t. } a(x, y) = \langle Ax, y \rangle \quad \forall y \in V. \end{aligned}$$

Remark 1.1. If A is a regularly dissipative operator, then A and A^* generate an analytic semigroup of pseudo-contractions; moreover there exists $\mathcal{K}_A \geq 1$ such that, for all $t > 0$, $\|e^{tA}\| \leq e^{\bar{\omega}t}$ and $\|Ae^{tA}\| \leq \mathcal{K}_A t^{-1} e^{\bar{\omega}t}$ (see [15]).

We set $C = \sum_{i=1}^{+\infty} \langle \cdot, e_i^{(1)} \rangle C_i$; $D = \sum_{i=1}^{+\infty} \langle \cdot, e_i^{(2)} \rangle D_i$, and we assume that the following hypotheses on the linear operators hold:

$$\begin{aligned}
 & \forall i \in \mathbb{N}, C_i \in \mathcal{L}(V, H) \quad \text{and} \quad \sum_{i=1}^{\infty} \lambda_i \|C_i\|_{\mathcal{L}(V, H)}^2 < +\infty; \\
 (\text{Hyp. 1}) \quad & \exists \eta \in]0, 1[, \exists \gamma \in \mathbb{R} \quad \text{s.t.} \quad \forall x \in V, \quad \sum_{i=1}^{+\infty} \lambda_i \|C_i x\|^2 \leq -2\eta a(x, x) + \gamma \|x\|^2; \\
 & \forall i \in \mathbb{N}, D_i \in \mathcal{L}(H) \quad \text{and} \quad \sum_{i=1}^{+\infty} \nu_i \|D_i\|^2 < \infty; \\
 & B \in \mathcal{L}(H); \quad P_0 \quad \text{and} \quad S \in \Sigma^+(H),
 \end{aligned}$$

where, by $\Sigma^+(H)$ (respectively, by $\Sigma(H)$) we denote the set of all self-adjoint and non-negative (respectively, self-adjoint) operators on H .

Moreover, we use the following hypotheses:

— K^1 and K^2 are two real separable Hilbert spaces, and $\{e_i^{(j)} : i \in \mathbb{N}\}$ is an orthonormal basis in K^j ($j = 1; 2$);

— $(\Omega, \mathcal{E}, \mathbf{P})$ is a complete probability space and $\{\mathcal{F}_t\}_{t \geq 0}$ is a filtration in it verifying the “usual hypotheses” (see [13]);

— $\{W_t^{(1)}\}_{t \geq 0}$ and $\{W_t^{(2)}\}_{t \geq 0}$ are two independent Wiener processes defined in Ω , adapted to the filtration \mathcal{F} and with values in K^1 and K^2 , respectively given by

$$W_t^{(1)} = \sum_{i=1}^{+\infty} \sqrt{\lambda_i} e_i^{(1)} \beta_i^{(1)}(t), \quad W_t^{(2)} = \sum_{i=1}^{+\infty} \sqrt{\nu_i} e_i^{(2)} \beta_i^{(2)}(t),$$

where $\{\beta_i^{(j)} : i \in \mathbb{N}\}$ ($j = 1; 2$) are two families of standard independent Brownian motions and $\sum_{i=1}^{+\infty} \lambda_i < +\infty$; $\sum_{i=1}^{+\infty} \nu_i < +\infty$.

If K is a Hilbert space, $M_{\mathcal{P}}^2(0, T, K)$ is the closed subspace of $L^2(\Omega \times [0, T], \mathcal{E} \otimes \mathcal{B}([0, T]), \mathbf{P} \otimes \mu, K)$ (where by μ we mean Lebesgue’s measure) given by all equivalence classes that contain a predictable process with respect to the filtration \mathcal{F} .

We denote by $L_{\mathcal{P}}^2(0, T, K)$ the set of all predictable processes $g : \Omega \times [0, T] \rightarrow K$ such that

$$\mathbf{P} \left\{ \int_0^T \|g(\sigma)\|^2 d\sigma < +\infty \right\} = 1.$$

Note that if for all $i \in \mathbb{N}$, $g_i \in M_{\mathcal{P}}^2(0, T, K)$, and if $\sum_{i=1}^{+\infty} \lambda_i \int_0^T \|g_i(\sigma)\|^2 d\sigma < +\infty$, then the sum $\sum_{i=1}^{+\infty} \sqrt{\lambda_i} \int_0^{\cdot} g_i(\sigma) d\beta^{(1)}(\sigma)$ converges in $L^2(\Omega, \mathcal{C}(0, T, K))$ (and the same holds true for the stochastic integration relative to $W^{(2)}$).

Now we return to problem $(S.E.)$; note that $(S.E.)$ can be solved only in a mild sense. On the other hand, as we have already noted, in the following we will need classical solutions of $(S.E.)$. Therefore we introduce below two different classes of approximating equations ($\{(S.E.)_h : h \in \mathbb{N}; h \geq \bar{\omega}\}$ and $\{(S.E.)_{h,k} : h, k \in \mathbb{N}; h, k \geq \bar{\omega}\}$, respectively:

$$\begin{aligned}
 (S.E.)_h \quad & dy_h = (Ay_h + Bu_h)dt + C_h y_h dW_t^{(1)} + Du_h dW_t^{(2)}, \\
 & y_h(0) = x_h;
 \end{aligned}$$

$$(S.E.)_{h,k} \quad \begin{aligned} dy_{h,k} &= (A_k y_{h,k} + B u_{h,k}) dt + C_h y dW_t^{(1)} + D u_{h,k} dW_t^{(2)}, \\ y_{h,k}(0) &= x_{h,k}, \end{aligned}$$

where $x_h, x_{h,k} \in L^2(\Omega, \mathcal{F}_0, P, H)$; $u_h, u_{h,k} \in M_P^2(0, T, H)$, and $A_k = AJ(k, A)$, $C_{i,h} = C_i J(h, A)$ (for all $\lambda \in \mathbb{C}$ subject to $\operatorname{Re} \lambda > \bar{\omega}$, we write $J(\lambda, A) = \lambda(\lambda I - A)^{-1}$). Note that, in general, only problem $(S.E.)_{h,k}$ has a classical solution.

DEFINITION. A process $y \in M_P^2(0, T, V)$ is a mild solution of $(S.E.)$ if, almost surely in $\Omega \times [0, T]$,

$$\begin{aligned} y(t) &= e^{tA} x + \int_0^t e^{(t-s)A} B u(s) ds + \sum_{i=1}^{+\infty} \lambda_i^{1/2} \int_0^t e^{(t-s)A} C_i y(s) d\beta_i^{(1)}(s) \\ &\quad + \sum_{i=1}^{+\infty} \nu_i^{1/2} \int_0^t e^{(t-s)A} D_i u(s) d\beta_i^{(2)}(s). \end{aligned}$$

DEFINITION. $y_h \in M_P^2(0, T, H)$ is called a mild solution of $(S.E.)_h$ if, almost surely in $\Omega \times [0, T]$,

$$\begin{aligned} y_h(t) &= e^{tA} x_h + \int_0^t e^{(t-s)A} B u_h(s) ds + \sum_{i=1}^{+\infty} \lambda_i^{1/2} \int_0^t e^{(t-s)A} C_{i,h} y_h(s) d\beta_i^{(1)}(s) \\ &\quad + \sum_{i=1}^{+\infty} \nu_i^{1/2} \int_0^t e^{(t-s)A} D_i u_h(s) d\beta_i^{(2)}(s). \end{aligned}$$

We have the following result.

THEOREM 1.1. Fix any $T > 0$; then the following statements hold:

- (i) There exists a unique mild solution y (respectively, y_h), belonging to $M_P^2(0, T, V)$, of $(S.E.)$ (respectively, $(S.E.)_h$).
- (ii) If $x_h \rightarrow x$ in $L^2(\Omega, \mathcal{F}_0, P, H)$ and $u_h \rightarrow u$ in $M_P^2(0, T, H)$, then $y_h \rightarrow y$ in $M_P^2(0, T, H)$.
- (iii) The equivalence class of y (respectively, y_h) in $L^2(\Omega \times [0, T], \mathcal{E} \otimes \mathcal{B}([0, T]), P \otimes \mu, H)$ contains a unique (up to a modification) stochastic process \tilde{y} (respectively, \tilde{y}_h) belonging to $\mathcal{C}(0, T, L^2(\Omega, \mathcal{E}, P, H))$. Moreover, if $x_h \rightarrow x$ in $L^2(\Omega, \mathcal{F}_0, P, H)$ and $u_h \rightarrow u$ in $M_P^2(0, T, H)$, then for all $t \in [0, T]$, $y_h(t) \rightarrow \tilde{y}(t)$ in $L^2(\Omega, \mathcal{F}_t, P, H)$ (in the following, by a mild solution of $(S.E.)$ (respectively, $(S.E.)_h$), we will always mean an element of $\mathcal{C}(0, T, L^2(\Omega, \mathcal{E}, P, H))$).
- (iv) For all $h, k \in \mathbb{N}$; $h, k \geq \bar{\omega}$ there exists a unique (up to a modification) mild solution $y_{h,k}$ (belonging to $M_P^2(0, T, H)$) of $(S.E.)_{h,k}$. Moreover, if for fixed h , $x_{h,k} \rightarrow x_h$ in $L^2(\Omega, \mathcal{F}_0, P, H)$ and $u_{h,k} \rightarrow u$ in $M_P^2(0, T, H)$, then $y_{h,k} \rightarrow y_h$ in $M_P^2(0, T, V)$ and for all $t \in [0, T]$, $y_{h,k}(t) \rightarrow \tilde{y}_h(t)$ in $L^2(\Omega, \mathcal{F}_t, P, H)$.

Proof. The argument that follows is due to Flandoli for what concerns point (i) and deals with some new difficulties in relation to point (ii).

Point (i). Let us define on V the following norm: $\|v\|_M^2 = -a(v, v) + M\|v\|^2$, where $M \geq \bar{\omega}$; clearly, $\|\cdot\|_M$ and $|\cdot|$ are equivalent for all $M \geq \bar{\omega}$.

In our hypotheses it is still possible to show (see [3]) that if $\{g_i : i \in \mathbb{N}\} \subset L_P^2(0, T, H)$ and $\sum_{i=1}^{\infty} \lambda_i \mathbb{E} \int_0^T \|g_i(\sigma)\|^2 d\sigma < +\infty$, then, for all $t \in [0, T]$, the following estimate

holds:

$$(1.1) \quad \int_0^T \mathbf{E} \left\| \left\| \sum_{i=1}^{\infty} \sqrt{\lambda_i} \int_0^t e^{(t-s)A} g_i(s) d\beta_i^{(1)}(s) \right\| \right\|_M^2 dt \\ \leq \left(\frac{1}{2} + T e^{2\bar{\omega}T} M \right) \sum_{i=1}^{\infty} \lambda_i \mathbf{E} \int_0^T \|g_i(s)\|^2 ds$$

(and an identical relation holds if $W^{(1)}$ is substituted by $W^{(2)}$). In fact,

$$\int_0^T \mathbf{E} \left(\left\| \left\| \sum_{i=1}^{\infty} \sqrt{\lambda_i} \int_0^t e^{(t-s)A} g_i(s) d\beta_i^{(1)}(s) \right\| \right\|_m^2 \right) dt \\ \leq \sum_{i=1}^{\infty} \lambda_i \mathbf{E} \int_0^T \left(\int_0^t \left(\langle -A e^{(t-s)A} g_i(s); e^{(t-s)A} g_i(s) \rangle + M \|e^{(t-s)A} g_i(s)\|^2 \right) ds \right) dt \\ \leq \sum_{i=1}^{\infty} \lambda_i \mathbf{E} \int_0^T \left(\int_0^{T-s} \langle -A e^{\sigma A} g_i(s); e^{\sigma A} g_i(s) \rangle d\sigma \right) ds \\ + T e^{2\bar{\omega}T} M \sum_{i=1}^{\infty} \lambda_i \int_0^T \mathbf{E} \|g_i(s)\|^2 ds \\ \leq \frac{1}{2} \sum_{i=1}^{\infty} \lambda_i \mathbf{E} \int_0^T \left(\int_0^T -\frac{d}{d\sigma} \|e^{\sigma A} g_i(s)\|_H^2 d\sigma \right) ds + T e^{2\bar{\omega}T} M \sum_{i=1}^{\infty} \lambda_i \int_0^T \mathbf{E} \|g_i(s)\|^2 ds \\ \leq \left(\frac{1}{2} + T e^{2\bar{\omega}T} M \right) \sum_{i=1}^{\infty} \lambda_i \int_0^T \mathbf{E} \|g_i(s)\|_H^2 ds.$$

Now let, for all $\xi \in \mathbf{M}_{\mathcal{P}}^2(0, T, V)$,

$$\chi(\xi)(t) = \sum_{i=1}^{+\infty} \lambda_i^{1/2} \int_0^t e^{(t-s)A} C_i \xi(s) d\beta_i^{(1)}(s),$$

let $\eta_1 \in]\eta, 1[$, and choose M and T_0 such that

$$(1.2) \quad M \geq \max \left\{ \frac{\gamma}{2\eta}, \bar{\omega} \right\} \quad \text{and} \quad T_0 \leq \min \left\{ 1, \frac{(\eta_1 - \eta)}{2M\eta} e^{-2\bar{\omega}} \right\}.$$

Then (if V is endowed with the norm $\|\cdot\|_M$) χ is a contraction in $\mathbf{M}_{\mathcal{P}}^2(0, T_0, V)$. In fact, by (1.1) and (Hyp. 1), we get

$$\int_0^{T_0} \mathbf{E} \|\chi(\xi)(t)\|_M^2 dt \leq \left(\frac{1}{2} + T_0 e^{2\bar{\omega}T_0} M \right) \int_0^{T_0} \mathbf{E} \sum_{i=1}^{\infty} \lambda_i \|C_i \xi(t)\|^2 dt \\ \leq 2\eta \left(\frac{1}{2} + T_0 e^{2\bar{\omega}T_0} M \right) \int_0^{T_0} \mathbf{E} \|\xi(t)\|_M^2 dt \\ \leq \eta_1 \mathbf{E} \int_0^{T_0} \|\xi(t)\|_M^2 dt.$$

Finally, for all $t \in [0, T]$,

$$F(t) = e^{tA} x + \int_0^t e^{(t-s)A} B u(s) + \sum_{i=1}^{+\infty} \nu_i^{1/2} \int_0^t e^{(t-s)A} D_i u(s) d\beta_i^{(2)}(s).$$

It is very easy to prove (using the above technique) that $F \in \mathbf{M}_{\mathcal{P}}^2(0, T, V)$. Therefore the local existence and uniqueness of the mild solution of $(S.E.)$ can be proved by a standard contraction argument in the space $\mathbf{M}_{\mathcal{P}}^2(0, T_0, V)$ (where V is endowed with the norm $||| \cdot |||_M$) and, since our equation is linear with time-constant coefficients, this automatically implies the existence and uniqueness of the mild solution in all $\mathbf{M}_{\mathcal{P}}^2(0, T, V)$ (for all fixed T).

Note that, since for all $v \in V$,

$$(1.3) \quad \sum_{i=1}^{\infty} \lambda_i \|C_i J(h, A)v\|^2 \leq 2\eta \langle -AJ(h, A)V, J(h, A)v \rangle + \gamma \|v\|^2 \\ \leq (2\eta \|A_h\|_{\mathcal{L}(H)} + \gamma) \|v\|^2,$$

the above argument enables us to prove, for every $(S.E.)_h$, the existence and uniqueness of a mild solution y_h that belongs not only to $\mathbf{M}_{\mathcal{P}}^2(0, T, H)$ but also to $\mathbf{M}_{\mathcal{P}}^2(0, T, V)$. However, since M and T_0 defined in (1.2) depend on γ we cannot prove (ii) simply by a parameter-dependent contraction argument.

Point (ii). Let, for all $h \in \mathbb{N}$, $\phi_h = J(h, A)y_h$ (where y_h is the mild solution of $(S.E.)_h$). Then, by definition, it holds (in $\mathbf{M}_{\mathcal{P}}^2(0, T, V)$) that

$$\phi_h(t) = e^{tA} J(h, A)x_h + \int_0^t e^{(t-s)A} J(h, A)Bu_h(s)ds \\ + \sum_{i=1}^{+\infty} \lambda_i^{1/2} \int_0^t e^{(t-s)A} J(h, A)C_i \phi_h(s) d\beta_i^{(1)}(s) \\ + \sum_{i=1}^{+\infty} \nu_i^{1/2} \int_0^t e^{(t-s)A} J(h, A)D_i u_h(s) d\beta_i^{(2)}(s).$$

Therefore if $\psi_h = y - \phi_h$, then

$$\psi_h = F_h + \chi_h(\psi_h),$$

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

$$F_h(t) = e^{tA}(x - J(h, A)x_h) + \sum_{i=1}^{+\infty} \nu_i^{1/2} \int_0^t e^{(t-s)A} (D_i u(s) - J(h, A)D_i u_h(s)) d\beta_i^{(2)}(s) \\ - \int_0^t e^{(t-s)A} (J(h, A)Bu_h(s) - Bu(s))ds \\ + \sum_{i=1}^{+\infty} \lambda_i^{1/2} \int_0^t e^{(t-s)A} (I - J(h, A))C_i y(s) d\beta_i^{(1)}(s), \\ \chi_h(\psi_h)(t) = \sum_{i=1}^{+\infty} \lambda_i^{1/2} \int_0^t e^{(t-s)A} J(h, A)C_i \psi_h(s) d\beta_i^{(1)}(s).$$

Using relation (1.1) and the fact that $\|J(h, A)\|_{\mathcal{L}(H)} \leq 1$, we can, arguing exactly as in point (i), prove that if T_0 and M verify (1.2) and if V is endowed with the norm $||| \cdot |||_M$, then χ is a contraction in $\mathbf{M}_{\mathcal{P}}^2(0, T_0, V)$. Moreover, from relation (1.1) it follows that if the hypotheses of point (ii) hold, then $F_h \rightarrow 0$ in the norm of $\mathbf{M}_{\mathcal{P}}^2(0, T_0, V)$ (as $h \rightarrow +\infty$).

We have therefore proved that $\psi_h \rightarrow 0$ in $\mathbf{M}_{\mathcal{P}}^2(0, T_0, V)$ (as $h \rightarrow +\infty$) and, since our problem is linear, repeating the same argument, we deduce that $\psi_h \rightarrow 0$ in $\mathbf{M}_{\mathcal{P}}^2(0, T, V)$. Finally, observe that

$$\begin{aligned} y(t) - y_h(t) &= e^{tA}(x - x_h) + \int_0^t e^{(t-s)A} B(u(s) - u_h(s)) ds \\ &\quad + \sum_{i=1}^{\infty} \lambda_i^{1/2} \int_0^t e^{(t-s)A} C_i \psi_h(s) d\beta_i^{(1)}(s) \\ &\quad + \sum_{i=1}^{\infty} \nu_i^{1/2} \int_0^t e^{(t-s)A} D_i(u(s) - u_h(s)) d\beta_i^{(2)}(s). \end{aligned}$$

Therefore if $x_h \rightarrow x$, $u_h \rightarrow u$ in $\mathbf{M}_{\mathcal{P}}^2(0, T, H)$ and $\psi_h \rightarrow 0$ in $\mathbf{M}_{\mathcal{P}}^2(0, T, H)$ (as we have proved this is true if the hypotheses of point (ii) hold), then $y_h \rightarrow y$ in $\mathbf{M}_{\mathcal{P}}^2(0, T, V)$.

Point (iii). Now let y be the mild solution of $(\mathcal{S.E.})$ and define \tilde{y} as follows:

$$\begin{aligned} \tilde{y}(t) &= e^{tA}x + \int_0^t e^{(t-s)A} Bu(s) ds + \sum_{i=1}^{+\infty} \lambda_i^{1/2} \int_0^t e^{(t-s)A} C_i y(s) d\beta_i^{(1)}(s) \\ (1.4) \quad &+ \sum_{i=1}^{+\infty} \nu_i^{1/2} \int_0^t e^{(t-s)A} D_i u(s) d\beta_i^{(2)}(s); \end{aligned}$$

the same definition is given for \tilde{y}_h .

From the definition of mild solution, it follows that \tilde{y} belongs to the equivalence class of y and that \tilde{y}_h belongs to the equivalence class of y_h . Moreover, it is very easy to prove that $\tilde{y} \in \mathcal{C}(0, T, \mathbf{L}^2(\Omega, \mathcal{E}, \mathbf{P}, H))$ and that it is unique up to a modification. All the other properties are easy consequences of (1.4) and of point (ii).

Point (iv). The existence of a classical solution $y_{h,k}$ of $(\mathcal{S.E.})_{h,k}$ follows from the remark that every $(\mathcal{S.E.})_{h,k}$ is equivalent to its mild form and from a straightforward contraction argument in $\mathbf{M}_{\mathcal{P}}^2(0, T, H)$. The convergence in $\mathbf{M}_{\mathcal{P}}^2(0, T, H)$ of $y_{h,k}$ toward y_h follows from a standard parameter-dependent contraction argument. The rest is identical to point (iii). \square

In the following sections of this work the next generalization will be useful.

Remark 1.2. Let $L, L_h, L_{h,k}, G, G_h, G_{h,k}$ (for all $h, k \in \mathbb{N}$; $h, k \geq \bar{\omega}$) be strongly continuous maps $[0, T] \rightarrow \mathcal{L}(H)$; then all the conclusions stated in Theorem 1.1 remain true if the $(\mathcal{S.E.})$ is replaced by the following closed loop equation:

$$\begin{aligned} (C.L.E.) \quad dy(s) &= (A + BL(s))y(s)ds + Cy(s)dW_s^{(1)} + DG(s)y(s)dW_s^{(2)}, \\ y(0) &= x; \end{aligned}$$

problems $(\mathcal{S.E.})_h$ and $(\mathcal{S.E.})_{h,k}$ are replaced by the following:

$$\begin{aligned} (C.L.E.)_h \quad dy_h(s) &= (A + BL_h(s))y_h(s)ds + C_h y_h(s)dW_s^{(1)} + DG_h(s)y_h(s)dW_s^{(2)}, \\ y_h(0) &= x_h; \end{aligned}$$

$$\begin{aligned} (C.L.E.)_{h,k} \quad dy_{h,k}(s) &= (A_k + BL_{h,k}(s))y_{h,k}(s)ds + C_h y_{h,k}(s)dW_s^{(1)} \\ &\quad + DG_{h,k}(s)y_{h,k}(s)dW_s^{(2)}, \\ y_{h,k}(0) &= x_{h,k}, \end{aligned}$$

and, finally, the hypotheses

$$“u_{h,k} \rightarrow u_h \text{ in } M_P^2(0, T, H)” \quad \text{and} \quad “u_h \rightarrow u \text{ in } M_P^2(0, T, H)”$$

are replaced by

$$“\forall x_0 \in H \quad L_{h,k}x_0 \rightarrow L_hx_0 \quad \text{and} \quad G_{h,k}x_0 \rightarrow G_hx_0 \text{ in } C(0, T, H)”$$

and

$$“\forall x_0 \in H \quad L_hx_0 \rightarrow Lx_0 \quad \text{and} \quad G_hx_0 \rightarrow Gx_0 \text{ in } C(0, T, H),”$$

respectively. Note that the terms containing L and G in the above equations are bounded perturbations of the previous ones and therefore do not add new difficulties.

2. Solution of the Riccati equation. We consider the Riccati equation

$$\begin{aligned} (\mathcal{R.E.}) \quad P' &= A^*P + PA + \sum_{i=1}^{+\infty} \lambda_i C_i^* P C_i + S - PB \left[I + \sum_{i=1}^{+\infty} \nu_i D_i^* P D_i \right]^{-1} B^* P, \\ P(0) &= P_0, \end{aligned}$$

as well as the following approximations (that correspond to problems $(\mathcal{S.E.})_h$ and $(\mathcal{S.E.})_{h,k}$ introduced in §1):

$$\begin{aligned} (\mathcal{R.E.})_h \quad P'_h &= A^*P_h + P_h A + \sum_{i=1}^{+\infty} \lambda_i C_{i,h}^* P_h C_{i,h} + S \\ &\quad - P_h B \left[I + \sum_{i=1}^{+\infty} \nu_i D_i^* P_h D_i \right]^{-1} B^* P_h, \\ P_h(0) &= P_{0,h}; \end{aligned}$$

$$\begin{aligned} (\mathcal{R.E.})_{h,k} \quad P'_{h,k} &= A_k^* P_{h,k} + P_{h,k} A_k + \sum_{i=1}^{+\infty} \lambda_i C_{i,h}^* P_{h,k} C_{i,h} + S \\ &\quad - P_{h,k} B \left[I + \sum_{i=1}^{+\infty} \nu_i D_i^* P_{h,k} D_i \right]^{-1} B^* P_{h,k}, \\ P_{h,k}(0) &= P_{0,h,k}, \end{aligned}$$

where $C_{i,h} = C_i J(h, A)$, $A_k = A J(k, A)$, and $P_{0,h}, P_{0,h,k} \in \Sigma^+(H)$.

Let us specify some notation. For all $\rho \in \mathbb{R}$, we define $\Sigma^+(H, \rho) = \{Q \in \Sigma(H) : Q \geq \rho\}$ (note that $\Sigma^+(H) = \Sigma^+(H, 0)$). By $\mathcal{C}_s(0, T, \Sigma(h))$ (respectively, $\mathcal{C}_s(0, T, \Sigma^+(H, \rho))$) we will denote the set of all strongly continuous mappings $P : [0, T] \rightarrow \Sigma(H)$ (respectively, $P : [0, T] \rightarrow \Sigma^+(H, \rho)$) (that is, of all mappings P such that $P(\cdot)x$ is continuous for all $x \in H$); this set will also be denoted by $\mathcal{C}_u(0, T, \Sigma(H))$ (respectively, $\mathcal{C}_u(0, T, \Sigma^+(H, \rho))$) when it is endowed with the “uniform convergence” norm

$$\|P\|_u = \sup_{t \in [0, T]} \{\|P(t)\|\}.$$

Let $\delta = \sum_{i=1}^{\infty} \nu_i \|D_i\|_{\mathcal{L}(H)}^2$.

DEFINITION. We say that $P \in \mathcal{C}_s(0, T, \Sigma^+, (H, -(2\delta)^{-1}))$ is a mild solution of $(\mathcal{R.E.})$ if, for all $x \in H$, $t > 0$,

$$\begin{aligned} P(t)x &= e^{tA^*} P_0 e^{tA} x + \sum_{i=1}^{+\infty} \lambda_i \int_0^t \left(C_i e^{(t-s)A} \right)^* P(s) \left(C_i e^{(t-s)A} \right) x ds \\ &\quad - \int_0^t e^{(t-s)A^*} P(s) B \left[I + \sum_{i=1}^{+\infty} \nu_i D_i^* P D_i \right]^{-1} B^* P(s) e^{(t-s)A} x ds \\ &\quad + \int_0^t e^{(t-s)A^*} S e^{(t-s)A} x ds. \end{aligned}$$

Similar definitions are given for the mild solutions of the approximating equations.

Observe that (Hyp. 1) yields $\sum_{i=1}^{\infty} \lambda_i \|C_{i,h}\|_{\mathcal{L}(H)}^2 < \infty$. Therefore each $(\mathcal{R.E.})_{h,k}$ is meaningful and, if $P_{h,k} \in \mathcal{C}^0([0, t], \Sigma^+(h, -(2\delta)^{-1}))$ verifies (2.1), where A is replaced by A_k and C_i by $C_{i,h}$, then $P_{h,k} \in \mathcal{C}^1([0, t], \Sigma^+(H, -(2\delta)^{-1}))$ and is a classical solution of $(\mathcal{R.E.})_{h,k}$; moreover, the mild form of every $(\mathcal{R.E.})_h$ makes sense.

Under (Hyp. 1) we have only $\sum_{i=1}^{\infty} \lambda_i \|C_i e^{(t-s)A} x\|^2 \leq \text{const } \|x\|^2 (t-s)^{-1}$; thus it is not a priori evident that expression (2.2) below is well defined:

$$(2.2) \quad \sum_{i=1}^{+\infty} \lambda_i \int_0^t \left(C_i e^{(t-s)A} \right)^* P(s) \left(C_i e^{(t-s)A} \right) x ds,$$

where $P \in \mathcal{C}_s(0, T, \Sigma(H))$.

Let us fix a positive time $T > 0$, and let us discuss any term of (2.1) starting from (2.2), the most difficult one.

LEMMA 2.1. *For every $P \in \mathcal{C}_s(0, T, \Sigma(H))$ and for every $x \in H$, $t \in [0, T]$ and $h, k \in \mathbb{N}$, we have*

$$(2.3) \quad \left| \sum_{i=1}^{+\infty} \lambda_i \int_0^t \left\langle x; \left(C_i e^{(t-s)A} \right)^* P(s) \left(C_i e^{(t-s)A} \right) x \right\rangle ds \right| \leq (\eta + t e^{2\bar{\omega}t} \gamma) \|P\|_u \|x\|^2,$$

$$(2.4) \quad \left| \sum_{i=1}^{+\infty} \lambda_i \int_0^t \left\langle x; \left(C_{i,h} e^{(t-s)A} \right)^* P(s) \left(C_{i,h} e^{(t-s)A} \right) x \right\rangle ds \right| \leq (\eta + t e^{2\bar{\omega}t} \gamma) \|P\|_u \|x\|^2,$$

$$(2.5) \quad \left| \sum_{i=1}^{+\infty} \lambda_i \int_0^t \left\langle x; \left(C_{i,h} e^{(t-s)A} \right)^* P(s) \left(C_{i,h} e^{(t-s)A} \right) x \right\rangle ds \right| \leq t e^{2\bar{\omega}t} (2\eta \|A_h\|_{\mathcal{L}(H)} + \gamma) \|P\|_u \|x\|^2,$$

$$(2.6) \quad \left| \sum_{i=1}^{+\infty} \lambda_i \int_0^t \left\langle x; \left(C_{i,h} e^{(t-s)A_k} \right)^* P(s) \left(C_{i,h} e^{(t-s)A_k} \right) x \right\rangle ds \right| \leq t e^{2\bar{\omega}t} (2\eta \|A_h\|_{\mathcal{L}(H)} + \gamma) \|P\|_u \|x\|^2,$$

Proof. We prove only (2.4), the proof of (2.3) being similar and (2.5), (2.6) being trivial consequences of (1.3). From (Hyp. 1) we have

$$\begin{aligned}
 & \left| \sum_{i=1}^{+\infty} \lambda_i \int_0^t \langle x; (C_{i,h} e^{(t-s)A})^* P(s) (C_{i,h} e^{(t-s)A}) x \rangle ds \right| \\
 & \leq 2\eta \|P\|_u \int_0^t \langle -A e^{\sigma A} J(h, A) x; e^{\sigma A} J(h, A) x \rangle d\sigma \\
 & \quad + \gamma \int_0^t \|P(s)\|_{\Sigma(H)} \|e^{\sigma A} J(h, A) x\|^2 d\sigma \\
 & \leq \eta \|P\|_u \int_0^t -\frac{d}{d\sigma} \|e^{\sigma A} J(h, A) x\|^2 d\sigma \\
 & \quad + \gamma e^{2\bar{\omega}t} \int_0^t \|P(s)\|_{\Sigma(H)} \|x\|^2 d\sigma \leq (\eta + t\gamma e^{2\bar{\omega}t}) \|P\|_u \|x\|^2,
 \end{aligned}
 \tag{2.7}$$

and this is exactly what we wanted to prove. \square

PROPOSITION 2.1. (i) *Let $x \in H$, $t \in [0, T]$; then there exists, in the sense of the strong convergence in H , the following limit:*

$$\Gamma(P)(t)x = \lim_{\epsilon \searrow 0, n \rightarrow +\infty} \sum_{i=1}^n \lambda_i \int_0^{t-\epsilon} (C_i e^{(t-s)A})^* P(s) (C_i e^{(t-s)A}) x ds.
 \tag{2.8}$$

- (ii) *Let $\Gamma_h(P)(t)x$ be defined as in (2.8), replacing C_i by $C_{i,h}$. Let $\{P_h : h \in \mathbb{N}\} \subset \mathcal{C}_s(0, T, \Sigma(H))$, and $P \in \mathcal{C}_s(0, T, \Sigma(H))$. Then if $P_h \rightarrow P$ strongly and uniformly in $[0, T]$, $\Gamma_h(P_h) \rightarrow \Gamma(P)$ strongly and uniformly in $[0, T]$.*
- (iii) *Let $\Gamma_{h,k}(P)(t)x$ be defined as in (2.8), replacing C_i by $C_{i,h}$ and A by A_k . Let, for fixed h , $\{P_{h,k} : h, k \in \mathbb{N}\} \subset \mathcal{C}_s(0, T, \Sigma(H))$, and $P_h \in \mathcal{C}_s(0, T, \Sigma(H))$. Then if $P_{h,k} \rightarrow P_h$ strongly and uniformly in $[0, T]$, then $\Gamma_{h,k}(P_{h,k}) \rightarrow \Gamma_h(P_h)$ strongly and uniformly in $[0, T]$.*

Proof. We prove points (i) and (ii) first. Let $P, P_h \in \mathcal{C}_s(0, T, \Sigma(H))$. Fix $M \geq \bar{\omega}$, suppose V endowed with the norm $\|\cdot\|_M$, and set $\rho = \sum_{i=1}^{\infty} \lambda_i \|C_i\|_{\mathcal{L}(V,H)}^2$. Then it holds that

$$\sum_{i=1}^{\infty} \lambda_i \|C_i e^{tA}\|_{\mathcal{L}(H)}^2 \leq \rho e^{2\bar{\omega}t} \left(\frac{\mathcal{K}_A}{t} + M \right),
 \tag{2.9}$$

and exactly the same holds true for $\sum_{i=1}^{\infty} \lambda_i \|C_{i,h} e^{tA}\|_{\mathcal{L}(H)}^2$ for all $h \in \mathbb{N}$, $h \geq \bar{\omega}$.

Step 1. Let us suppose that $x \in \mathcal{D}(A)$. Then

$$\begin{aligned}
 \sum_{i=1}^{+\infty} \lambda_i \|C_i e^{tA} x\|^2 & \leq -2\eta \langle e^{tA} A x, e^{tA} x \rangle + \gamma \|e^{tA} x\|^2 \\
 & \leq e^{2\bar{\omega}t} (2\eta \|x\| \|Ax\| + \gamma \|x\|^2),
 \end{aligned}
 \tag{2.10}$$

and exactly the same holds true (if $x \in \mathcal{D}(A)$) for $\sum_{i=1}^{\infty} \lambda_i \|C_{i,h} e^{tA} x\|^2$, for all $h \in \mathbb{N}$. Observe now that

$$\begin{aligned}
 & \sum_{i=1}^{\infty} \lambda_i \left\| (C_i e^{(t-s)A})^* P(s) (C_i e^{(t-s)A}) x \right\| \\
 & \leq \|P\|_u e^{2\bar{\omega}t} [\rho(M + \mathcal{K}_A/(t-s))(2\eta \|Ax\| \|x\| + \gamma \|x\|^2)]^{1/2},
 \end{aligned}$$

and the same estimate holds for $\sum_{i=1}^{\infty} \lambda_i \left\| (C_{i,h} e^{(t-s)A})^* P_h(s) (C_{i,h} e^{(t-s)A}) x \right\|$. Therefore we can define

$$(2.11) \quad \begin{aligned} \Gamma(P)(t)x &= \sum_{i=1}^{\infty} \lambda_i \int_0^t (C_i e^{(t-s)A})^* P(s) (C_i e^{(t-s)A}) x ds \\ &= \lim_{\epsilon \searrow 0, n \rightarrow +\infty} \sum_{i=1}^n \lambda_i \int_0^{t-\epsilon} (C_i e^{(t-s)A})^* P(s) (C_i e^{(t-s)A}) x ds, \end{aligned}$$

$$(2.12) \quad \begin{aligned} \Gamma_h(P)(t)x &= \sum_{i=1}^{\infty} \lambda_i \int_0^t (C_i e^{(t-s)A})^* P_h(s) (C_i e^{(t-s)A}) x ds \\ &= \lim_{\epsilon \searrow 0, n \rightarrow +\infty} \sum_{i=1}^n \lambda_i \int_0^{t-\epsilon} (C_{i,h} e^{(t-s)A})^* P_h(s) (C_{i,h} e^{(t-s)A}) x ds. \end{aligned}$$

Moreover, by the dominated convergence theorem, we deduce that the convergence of the left-hand side of (2.11) is uniform in $t \in [0, T]$ and, if the hypothesis of point (ii) holds, the convergence of the left-hand side of (2.12) is uniform in $t \in [0, T]$ and $h \in \mathbb{N}$. Thus, if $x \in \mathcal{D}(A)$, point (i) is proved. Moreover, again applying the dominated convergence theorem, we can conclude that point (ii) is proved (in the same particular case) if we show that, for all $i \in \mathbb{N}$, $\sigma > 0$,

$$\sup_{t \in [0, T]} \left\| (C_{i,h} e^{\sigma A})^* P_h(t) C_{i,h} e^{\sigma A} x - (C_i e^{\sigma A})^* P(t) C_i e^{\sigma A} x \right\| \rightarrow 0,$$

which is an easy consequence of well-known properties of the Yoshida approximations (note that for all $x \in H$, $(C_{i,h} e^{tA})^* x = J(h, A^*) (C_i e^{tA})^* x \rightarrow (C_i e^{tA})^* x$).

Step 2. To end the proof of points (i) and (ii), it is enough, taking into account Step 1 and the density of $\mathcal{D}(A)$ into H , to show that if the sequence $\{\|P_h\|_u; h \in \mathbb{N}\}$ is bounded, then the maps

$$\begin{aligned} x &\rightarrow \sum_{i=1}^N \lambda_i \int_0^{t-\epsilon} \langle C_i e^{(t-s)A} y; P(s) C_i e^{(t-s)A} x \rangle ds, \\ x &\rightarrow \sum_{i=1}^N \lambda_i \int_0^{t-\epsilon} \langle C_{i,h} e^{(t-s)A} y; P_h(s) C_{i,h} e^{(t-s)A} x \rangle ds \end{aligned}$$

are uniformly bounded in t , N , ϵ , y (if we suppose $\|y\| \leq 1$), and h . This is a trivial consequence of Lemma 2.1, however.

Finally, (iii) follows from standard calculations. \square

Remark 2.1. Since it is immediate to verify that $\Gamma_h(P)$ and $\Gamma_{h,k}(P)$ belong to $\mathcal{C}_s(0, T, \Sigma(H))$ (for all $P \in \mathcal{C}_s(0, T, \Sigma(H))$; for all $h, k \in \mathbb{N}$; $h, k, \geq \bar{\omega}$), by Proposition 2.1 we deduce that, for all $P \in \mathcal{C}_s(0, T, \Sigma(H))$, $\Gamma(P)$ belongs to $\mathcal{C}_s(0, T, \Sigma(H))$.

Let us examine the next term. We define

$$\mathbf{B}(0, T, R, -(2\delta)^{-1}) = \{P \in \mathcal{C}_s(0, T, \Sigma^+(H, (-2\delta)^{-1})) \text{ s.t. } \|P\|_u \leq R\}.$$

Given $P \in \mathcal{C}_s(0, T, \Sigma^+(H, (-2\delta)^{-1}))$, we set

$$\begin{aligned} \Psi(P)(t)x &= - \int_0^t e^{(t-s)A} P(s) B \left[I + \sum_{i=1}^{+\infty} \nu_i D_i^* P(s) D_i \right]^{-1} B^* P(s) e^{(t-s)A} x ds, \\ \Psi_k(P)(t)x &= - \int_0^t e^{(t-s)A} P(s) B \left[I + \sum_{i=1}^{+\infty} \nu_i D_i^* P(s) D_i \right]^{-1} B^* P(s) e^{(t-s)A} x ds. \end{aligned}$$

The following statements can be easily checked:

- $\Psi(P), \Psi_k(P) \in \mathcal{C}_s(0, T, \Sigma(H))$; $\Psi(P)(t), \Psi_k(P)(t) \leq 0$ (for all $t \in [0, T]$).
- There exists a positive constant \mathcal{K}_Ψ such that, for all $R > 0$ and P ,

$$\bar{P} \in \mathbf{B}(0, T, R, -(2\delta)^{-1}),$$

$$(2.13) \quad \|\Psi(P)\|_u \leq \mathcal{K}_\Psi T e^{2\bar{\omega}T} R^2; \quad \|\Psi_k(P)\|_u \leq \mathcal{K}_\Psi T e^{2\bar{\omega}T} R^2,$$

$$(2.14) \quad \|\Psi(P) - \Psi(\bar{P})\|_u \leq T e^{2\bar{\omega}T} \mathcal{K}_\Psi (R + R^2) \|P - \bar{P}\|_u,$$

$$(2.15) \quad \|\Psi_k(P) - \Psi_k(\bar{P})\|_u \leq T \mathcal{K}_\Psi e^{2\bar{\omega}T} (R + R^2) \|P - \bar{P}\|_u.$$

—If, for all $k \in \mathbb{N}$, $P_k \in \mathcal{C}_s(0, T, \Sigma^+(H, -(2\delta)^{-1}))$ and if $P_k \rightarrow P$ strongly and uniformly in $[0, T]$, then $\Psi_k(P_k) \rightarrow \Psi(P)$ strongly and uniformly in $[0, T]$.

Finally let us set, for any given $Q \in \Sigma(H)$,

$$\begin{aligned} F(Q)(t)x &= e^{tA^*} Q e^{tA} x + \int_0^t e^{(t-s)A^*} S e^{(t-s)A} x \, ds, \\ F_k(Q)(t)x &= e^{tA_k^*} Q e^{tA_k} x + \int_0^t e^{(t-s)A_k^*} S e^{(t-s)A_k} x \, ds. \end{aligned}$$

Then it is almost immediate to see that

— $F(Q), F_k(Q) \in \mathcal{C}_s(0, T, \Sigma(H))$ if $Q \in \Sigma^+(H)$, then $F(Q), F_k(Q) \in \mathcal{C}_s(0, T, \Sigma^+(H))$,

$$(2.16) \quad \|F(Q)\|_u \leq e^{2\bar{\omega}T} (\|Q\| + T\|S\|); \quad \|F_k(Q)\|_u \leq e^{2\bar{\omega}T} (\|Q\| + T\|S\|),$$

—If, for all $k \in \mathbb{N}$, $Q_k \in \Sigma^+(H)$ and if $Q_k \rightarrow Q$ strongly, then $F_k(Q_k) \rightarrow F(Q)$ strongly and uniformly in $[0, T]$.

We are now in a position to prove the local existence of a mild solution of $(\mathcal{R.E.})$. Set $\Xi = \Gamma + \Psi$; $\Xi_h = \Gamma_h + \Psi$; $\Xi_{h,k} = \Gamma_{h,k} + \Psi_k$.

Then $P \in \mathcal{C}_s(0, T, \Sigma^+(H, -(2\delta)^{-1}))$ is a mild solution of $(\mathcal{R.E.})$ if and only if

$$P(t)x = F(P_0)(t)x + \Xi(P)(t)x \quad \forall t \geq 0; x \in H,$$

and similar expressions can be given for the mild solution of $(\mathcal{R.E.})_h$ and $(\mathcal{R.E.})_{h,k}$.

Note that Γ, Γ_h , and $\Gamma_{h,k}$ are monotone (by this we mean that if $P \geq 0$, then $\Gamma(P) \geq 0$); therefore if P belongs to $\mathcal{C}_s(0, T, \Sigma^+(H, -(2\delta)^{-1}))$, the following estimates hold:

$$(2.17) \quad \Gamma(P) \geq \Gamma \left(-\frac{1}{2\delta} \right) \geq -\frac{\eta}{2\delta} - \frac{\gamma}{2\delta} T e^{2\bar{\omega}T},$$

$$(2.18) \quad \Gamma_{h,k}(P) \geq -\frac{2\|A_h\|_{\mathcal{L}(H)} + \gamma}{2\delta} T e^{2\bar{\omega}T} \quad \forall h, k \in \mathbb{N},$$

and both the previous estimates hold for Γ_h .

We can now prove the local existence of the solutions.

PROPOSITION 2.2. (i) Suppose that $0 \leq P_0, P_{0,h} \leq R$ (for all natural number $h \geq \bar{\omega}$); then there exists $T_0 > 0$ such that $\exists! P \in \mathcal{C}_s(0, T_0, \Sigma^+(H, -(2\delta)^{-1}))$ (respectively, $\exists! P_h \in \mathcal{C}_s(0, T_0, \Sigma^+(H, -(2\delta)^{-1}))$) mild solution of $(\mathcal{R.E.})$ (respectively, mild solution of $(\mathcal{R.E.})_h$). Moreover, if $P_{0,h} \rightarrow P_0$ strongly, then $P_h \rightarrow P$ strongly and uniformly in $[0, T]$.

(ii) Fix $h \in \mathbb{N}$ and suppose that $0 \leq P_{0,h}, P_{0,h,k} \leq R$ (for all $k \in \mathbb{N}$), then there exists $T_{0,h} > 0$ such that $\exists! P_h \in \mathcal{C}_s(0, T_0, \Sigma^+(H, -(2\delta)^{-1}))$ (respectively, $\exists! P_{h,k} \in \mathcal{C}^0(0, T_0, \Sigma^+(H, -(2\delta)^{-1}))$) mild solution of $(\mathcal{R.E.})_h$ (respectively, classical solution of $(\mathcal{R.E.})_{h,k}$). Moreover, if $P_{0,h,k} \rightarrow P_{0,h}$ strongly, then $P_{h,k} \rightarrow P_h$ strongly and uniformly in $[0, T_{0,h}]$.

Proof. First, we prove point (i). If we show that there exists $T_0 > 0, R_0 > 0, \eta_1 \in]0, 1[$ such that

$$\Xi(\mathbf{B}(0, T_0, R_0, -(2\delta)^{-1})) + F(P_0) \subset \mathbf{B}(0, T_0, R_0, -(2\delta)^{-1}),$$

$$\Xi_h(\mathbf{B}(0, T_0, R_0, -(2\delta)^{-1})) + F(P_{0,h}) \subset \mathbf{B}(0, T_0, R_0, -(2\delta)^{-1}),$$

and for all $P, \bar{P} \in \mathbf{B}(0, T_0, R_0, -(2\delta)^{-1})$,

$$\|\Xi(P) - \Xi(\bar{P})\|_{\mathcal{C}_u(0, T_0, \Sigma^+(H, -(2\delta)^{-1}))} \leq \eta_1 \|P - \bar{P}\|_{\mathcal{C}_u(0, T_0, \Sigma^+(H, -(2\delta)^{-1}))},$$

$$\|\Xi_h(P) - \Xi_h(\bar{P})\|_{\mathcal{C}_u(0, T_0, \Sigma^+(H, -(2\delta)^{-1}))} \leq \eta_1 \|P - \bar{P}\|_{\mathcal{C}_u(0, T_0, \Sigma^+(H, -(2\delta)^{-1}))},$$

then the claim (of point (i)) follows from a standard parameter-dependent fixed-point argument. From (2.13) and (2.17) we deduce that, for all $P \in \mathbf{B}(0, T_0, R_0, -(2\delta)^{-1})$, $t \in [0, t_0]$;

$$\Xi(P)(t) \geq - \left[\mathcal{K}_\Psi T_0 e^{2\bar{\omega} T_0} R_0^2 + \frac{\eta}{2\delta} + \frac{\gamma T_0}{2\delta} e^{2\bar{\omega} T_0} \right];$$

from (2.3), (2.13), and (2.16) we deduce that, for all $P \in \mathbf{B}(0, T_0, R_0, -(2\delta)^{-1})$,

$$\|\Xi(P) + F(P_0)\|_{\mathcal{C}_u(0, T_0, \Sigma(H))} \leq \eta R_0 + e^{2\bar{\omega} T_0} (\gamma T_0 R_0 + \mathcal{K}_\Psi T_0 R_0^2 + R + T_0 \|S\|);$$

and from (2.3) and (2.14) we deduce that, for all $P, \bar{P} \in \mathbf{B}(0, T_0, R_0, -(2\delta)^{-1})$,

$$\|\Xi(P) - \Xi(\bar{P})\|_{\mathcal{C}_u(0, T_0, \Sigma(H))} \leq (\eta + \gamma T_0 e^{2\bar{\omega} T_0} + \mathcal{K}_\Psi T_0 e^{2\bar{\omega} T_0} (R_0 + R_0^2)) \|P - \bar{P}\|_{\mathcal{C}_u(0, T_0, \Sigma(H))}.$$

So it is enough to choose $R_0 > e^{2\bar{\omega}} R(1 - \eta)^{-1}$, $\eta_1 \in]\eta, 1[$, and

$$T_0 < \min \left\{ 1, \frac{(1 - \eta)e^{-2\bar{\omega}}}{2\delta \mathcal{K}_\Psi R_0^2 + \gamma}; \frac{R_0 e^{-2\bar{\omega}}(1 - \eta) - R}{\mathcal{K}_\Psi R_0^2 + \gamma R_0 + \|S\|}; \frac{(\eta_1 - \eta)e^{-2\bar{\omega}}}{\mathcal{K}_\Psi (R_0 + R_0^2) + \gamma} \right\}.$$

Finally, let us observe, in relation to point (ii), that for all $Q \in \Sigma^+(H)$ and for all V in $\mathcal{C}^0(0, T, \Sigma^+(H, -(2\delta)^{-1}))$, $\Xi_{h,k}(V) + F_{h,k}(Q)$ belongs to $\mathcal{C}^0(0, T, \Sigma(H))$. Hence, the above-mentioned fixed-point argument can be carried on in $\mathcal{C}^0(0, T, \Sigma^+(H, -(2\delta)^{-1}))$, and therefore the function that we find this way is a classical solution of $(\mathcal{R.E.})_{h,k}$. Otherwise, the proof is identical to the one of point (i). \square

Through several applications of the previous proposition we get the following corollary.

COROLLARY 2.1. Suppose that there exists $P \in \mathcal{C}_s(0, \tau, \Sigma^+(H, -(2\delta)^{-1}))$ mild solution of $(\mathcal{R.E.})$ (respectively, $P_h \in \mathcal{C}_s(0, \tau, \Sigma^+(H, -(2\delta)^{-1}))$ mild solution of $(\mathcal{R.E.})_h$; respectively, $P_{h,k} \in \mathcal{C}^0(0, \tau, \Sigma^+(H, -(2\delta)^{-1}))$ solution of $(\mathcal{R.E.})_{h,k}$). Then P (respectively, P_h ; respectively, $P_{h,k}$) is unique in all $[0, \tau]$. Moreover, if $P_{0,h} \rightarrow P_0$ (respectively, fixed h , $P_{0,h,k} \rightarrow P_{0,h}$) strongly, then $P_h \rightarrow P$ (respectively, $P_{h,k} \rightarrow P_h$) strongly and uniformly in all $[0, \tau]$.

We now want to iterate the previous argument in $[T_{\xi,R}, 2T_{\xi,R}]$, and so on. To do this, we will show that the local solutions that we have just found are positive and fulfill an “a priori” bound.

PROPOSITION 2.3. *Let $G \in \mathcal{C}_s(0, \tau, \Sigma^+(H, -(2\delta)^{-1}))$ be a mild solution of $(\mathcal{R.E.})$ (respectively, a mild solution of $(\mathcal{R.E.})_h$; respectively, a solution of $(\mathcal{R.E.})_{h,k}$). Then, for all $t \in [0, \tau]$, we have $G(t) \geq 0$.*

Proof. By Corollary 2.1, it is enough to prove the claim when $G \in \mathcal{C}^1(0, T, \Sigma^+(H, -(2\delta)^{-1}))$ and is a classical solution of

$$(2.19) \quad \begin{aligned} G' &= A_k^* G + G A_k + \sum_{i=1}^{+\infty} \lambda_i C_{i,h}^* G C_{i,h} + S - G B \left[I + \sum_{i=1}^{+\infty} \nu_i D_i^* G D_i \right]^{-1} B^* G, \\ G(0) &= G_0, \end{aligned}$$

for some $h, k \in \mathbb{N}$, $h, k \geq \bar{\omega}$, and some $G_0 \geq 0$.

The proof follows in a standard way. Let $L = A_k - \frac{1}{2}B \left[I + \sum_{i=1}^{+\infty} \nu_i D_i^* G D_i \right]^{-1} B^* G$; clearly, $L \in \mathcal{C}^0(0, \tau, \mathcal{L}(H))$ and (2.19) can be written as

$$(2.20) \quad \begin{aligned} G' &= H L^* G + G L + \sum_{i=1}^{+\infty} \lambda_i C_{i,h}^* G C_{i,h} + S, \\ G(0) &= G_0. \end{aligned}$$

Now we observe that (2.20) is equivalent to

$$(2.21) \quad \begin{aligned} G(t) &= U(0, t) G_0 U^*(0, t) + \sum_{i=1}^{+\infty} \lambda_i \int_0^t U(s, t) C_{i,h}^* G(s) C_{i,h} U^*(s, t) ds \\ &\quad + \int_0^t U(s, t) S U^*(s, t) ds, \end{aligned}$$

where $U(t, s)$ is the evolution operator associated to $\{L^*(t)\}_{t \in [0, \tau]}$. Now we can solve (2.21) by successive approximations, and it is easy to see that all these approximations have positive values. Then the conclusion follows. \square

PROPOSITION 2.4. *Fix $\beta > 0$ and $T > 0$; there exists a positive constant K_0 such that if, for some $\tau \in [0, T]$, $G \in \mathcal{C}_s(0, \tau, \Sigma^+(H))$ is a mild solution of $(\mathcal{R.E.})$ (or a mild solution of $(\mathcal{R.E.})_n$ for some $n \in \mathbb{N}$, $n \geq \bar{\omega}$) verifying $0 \leq G(0) \leq \beta$, then $G(t) \leq K_0 \quad \forall t \in [0, \tau]$. Moreover, for all $h \in \mathbb{N}$, $h \geq \bar{\omega}$, there exists a positive constant $K_{0,h}$ such that if, for some $\tau \in [0, T]$, $G \in \mathcal{C}_s(0, \tau, \Sigma^+(H))$ is a mild solution of $(\mathcal{R.E.})_h$ (or a classical solution of $(\mathcal{R.E.})_{h,k}$ (for some $k \in \mathbb{N}$, $k \geq \bar{\omega}$) verifying $0 \leq G(0) \leq \beta$, then $G(t) \leq K_{0,h} \quad \forall t \in [0, \tau]$.*

Proof. Let $G \in \mathcal{C}_s(0, T, \Sigma^+(H))$ be a mild solution of $(\mathcal{R.E.})$. We can write

$$G(t)x = F(G_0)(t)x + \Xi(G)(t)x.$$

Then by (2.7) and (2.16) it follows that, for all $t \in [0, \tau]$,

$$\|G(t)\| \leq \bar{\beta} + \eta \|G\|_{\mathcal{C}_u(0, t, \Sigma^+(H))} + \gamma e^{2\bar{\omega}T} \int_0^t \|G(s)\| ds,$$

where $\bar{\beta} = e^{2\bar{\omega}T}(\beta + T\|S\|)$. Set $\theta(t) = \|G\|_{\mathcal{C}_u(0, t, \Sigma^+(H))}$. Then we have

$$(1 - \eta)\theta(t) \leq \bar{\beta} + \gamma e^{2\bar{\omega}T} \int_0^t \theta(s) ds \quad \forall t \in [0, \tau],$$

and, by the Gronwall lemma, we can conclude that there exists a positive constant \mathcal{K}_0 depending only from T , η , and $\bar{\beta}$ such that

$$\|G\|_{C_u(0,\tau,\Sigma^+(H))} \leq \mathcal{K}_0.$$

The rest of the proof is identical. \square

Resuming, we have the following theorem.

THEOREM 2.1. *Fix $T > 0$. Then*

- (i) *There exists a unique mild solution P (respectively, P_h) belonging to $C_s(0, T, \Sigma^+(H))$ of $(\mathcal{R.E.})$ (respectively, $(\mathcal{R.E.})_h$).*
- (ii) *For all $h, k \in \mathbb{N}$, there exists a unique classical solution $P_{h,k} \in C^1(0, T, \Sigma^+(h))$ of $(\mathcal{R.E.})_{h,k}$.*
- (iii) *If, for fixed h , $P_{0,h,k} \rightarrow P_{0,h}$ strongly as $k \rightarrow +\infty$ (respectively, if $P_{0,h} \rightarrow P_0$ strongly as $h \rightarrow +\infty$), then $P_{h,k} \rightarrow P_h$ strongly and uniformly in $[0, T]$ (respectively, $P_h \rightarrow P$ strongly and uniformly in $[0, T]$).*

We now want to show a monotonicity property for the mild solutions of $(\mathcal{R.E.})$.

THEOREM 2.2. *Let $G_j \in C_s(0, T, \Sigma^+(H))$ ($j = 1; 2$) be the mild solution of*

$$G'_j = A^*G_j + G_jA + \sum_{i=1}^{+\infty} \lambda_i C_i^* G_j C_i + S_j - G_j B \left[I + \sum_{i=1}^{+\infty} \nu_i D_i^* G_j D_i \right]^{-1} B^* G_j,$$

$$G_j(0) = G_{0,j}$$

and assume that $G_{1,0} \geq G_{2,0} \geq 0$ and $S_1 \geq S_2 \geq 0$; then we have $G_1(t) \geq G_2(t)$ for all $t \in [0, T]$. The same is true for any of the approximating equations.

Proof. By Theorem 2.1 it is enough to prove the claim when $G_j \in C^1(0, T, \Sigma^+(H))$ ($j = 1; 2$) and is a classical solution of the problem

$$G'_j = \Pi(G_j) + S_j - G_j B \left[I + \sum_{i=1}^{+\infty} \nu_i D_i^* G_j D_i \right]^{-1} B^* G_j,$$

$$G_j(0) = G_{0,j},$$

where $\Pi(Q) = A^*Q + QA_k + \sum_{i=1}^{+\infty} \lambda_i C_{i,h}^* Q C_{i,h}$ for some $h, k \in \mathbb{N}$ and $G_{1,0} \geq G_{2,0} \geq 0$, $S_1 \geq S_2 \geq 0$. Moreover, it is enough to prove that $G_1 \geq G_2$ in a neighbourhood of 0. Let $G_j^0 = 0$ ($j = 1; 2$) and let G_j^n ($j = 1; 2, n \in \mathbb{N}$) be the solution of

$$(G_j^n)' = \Pi(G_j^n) + S_j - G_j^n B \left[I + \sum_{i=1}^{+\infty} \nu_i D_i^* G_j^{n-1} D_i \right]^{-1} B^* G_j^n,$$

$$G_j(0) = G_{0,j}^n.$$

The existence uniqueness and uniform boundedness of the G_j^n 's (in the space $C^0(0, T, \Sigma^+(H))$) can be proved, as in theorem 2.1. Let, for all $Q \in \Sigma^+(H)$,

$$(2.22) \quad \mathcal{M}(Q) = \left[I + \sum_{i=1}^{+\infty} \nu_i D_i^* Q D_i \right]^{-1}$$

By the Lipschitz continuity (in $\Sigma^+(H)$) of the function $Q \rightarrow \mathcal{M}(Q)$, we deduce that if \mathcal{K}_1 and \mathcal{K}_2 are large enough,

$$\|G_j^{n+1} - G_j^n\|_{C_u(0,\tau,\Sigma(H))} \leq \tau \mathcal{K}_1 \|G_j^{n+1} - G_j^n\|_{C_u(0,\tau,\Sigma(H))} + \tau \mathcal{K}_2 \|G_j^n - G_j^{n-1}\|_{C_u(0,\tau,\Sigma(H))}.$$

So if τ is sufficiently small, we have that

$$\|G_j^{n+1} - G_j^n\|_{\mathcal{C}_u(0,\tau,\Sigma(H))} \leq \frac{1}{2} \|G_j^n - G_j^{n-1}\|_{\mathcal{C}_u(0,\tau,\Sigma(H))},$$

and, consequently, that $G_j^n \rightarrow G_j$ in $\mathcal{C}^0(0, \tau, \Sigma(H))$. The proof is complete if we show that $G_1^n \geq G_2^n$ for all $n \in \mathbb{N}$; this is done, in the following, by an induction argument.

The above relation is true for $n = 0$; suppose that it holds for $n-1$. Let $R = G_1^n - G_2^n$, $S_0 = S_1 - S_2$, $G_0 = G_{1,0} - G_{2,0}$; then R verifies

$$\begin{aligned} R' &= \Pi(R) + S_0 - G_1^n(t)B\mathcal{M}(G_1^{n-1})B^*G_1^n(t) + G_2^n(t)B\mathcal{M}(G_2^{n-1})B^*G_2^n(t), \\ R(0) &= R_0, \end{aligned}$$

or, equivalently,

$$\begin{aligned} R'(t) &= L^*(t)R(t) + R(t)L(t) + \sum_{i=1}^{+\infty} \lambda_i C_i^* R(t) C_i + \tilde{S}(t), \\ R(0) &= R(0), \end{aligned}$$

where

$$L(t) = A_k - \frac{1}{2} B\mathcal{M}(G_1^{n-1}(t))B^*(G_1^n(t) + G_2^n(t)),$$

$$\tilde{S}(t) = S_0 + G_2^n B [\mathcal{M}(G_2^{n-1}(t)) - \mathcal{M}(G_1^{n-1}(t))] B^* G_2^n.$$

Let us now observe that $L \in \mathcal{C}^0(0, \tau, \mathcal{L}(H))$ and that the monotonicity of the map $Q \rightarrow \mathcal{M}(Q)$ yields $\tilde{S} \in \mathcal{C}^0(0, \tau, \Sigma^+(H))$. Therefore, arguing exactly as we did in the proof of Proposition 2.3, we can see that $R(t) \geq 0$ for all $t \in [0, \tau]$. \square

COROLLARY 2.2. *Let \underline{P} be the mild solution of $(\mathcal{R.E.})$ with initial data $\underline{P}(0) = 0$. Then we have $\underline{P}(t+h) \geq \underline{P}(t)$ (for all $t \geq 0, h \geq 0$).*

Proof. Fix $h \geq 0$ and let P_h be the mild solution of $(\mathcal{R.E.})$ with $P_h(0) = \underline{P}(h)$. Then Theorem 2.2 yields $P_h(t) \geq \underline{P}(t)$ for all $t \geq 0$ and, by the uniqueness of the mild solution of $(\mathcal{R.E.})$, we deduce immediately that $P_h(t) = \underline{P}(t+h)$ (for all $t \geq 0, h \geq 0$). This completes the proof. \square

3. Dynamic programming and synthesis of the optimal control (finite horizon case). The following result is proved in a standard way; however, we give a complete proof, since some technical points require a careful analysis.

THEOREM 3.1. *Let $P \in \mathcal{C}_s(0, T, \Sigma^+(H))$ be the mild solution of $(\mathcal{R.E.})$ and let $y \in \mathbf{M}_{\mathcal{P}}^2(0, T, H)$ be the mild solution of $(\mathcal{S.E.})$. Then, for all $t \geq 0$,*

$$\begin{aligned} \mathbf{E}\langle P(t)x, x \rangle - \mathbf{E}\langle P_0 y(t), y(t) \rangle &= \mathbf{E} \int_0^t \left\| \sqrt{S}y(s) \right\|^2 + \|u(s)\|^2 ds \\ &\quad - \mathbf{E} \int_0^t \left\| \left(I + \sum_{i=1}^{+\infty} \nu_i D_i^* P(t-s) d_i \right)^{\frac{1}{2}} [\mathcal{M}(P(t-s))B^*P(t-s)y(s) + u(s)] \right\|^2 ds \end{aligned} \quad (3.1)$$

(for the definition of \mathcal{M} , see (2.22)).

Proof. Due to Theorems 1.1 and 2.1, it is sufficient to prove (3.1), where P is replaced by $P_{h,k}$ and y is replaced by $y_{h,k}$ (where $P_{h,k}$ is the solution of $(\mathcal{R.E.})_{h,k}$ and $y_{h,k}$ is

the solution of $(\mathcal{S.E.})_{h,k}$. Fix $t > 0$ and let $\phi_i(s) = \langle P_{h,k}(t-s)C_{i,h}y_{h,k}(s), y_{h,k}(s) \rangle$, $\psi_i(s) = \langle P_{h,k}(t-s)D_i u(s), y_{h,k}(s) \rangle$. By the Itô formula, we get

$$\begin{aligned} d_s \langle P_{h,k}(t-s)y_{h,k}(s), y_{h,k}(s) \rangle &= \sum_{i=1}^{+\infty} \sqrt{\lambda_i} \phi_i(s) d\beta_i^{(1)}(s) + \sum_{j=1}^{+\infty} \sqrt{\nu_j} \psi_j(s) d\beta_j^{(2)}(s) \\ &+ \left\{ \left\| \left(I + \sum_{i=1}^{+\infty} \nu_i D_i^* P_{h,k}(t-s) D_i \right)^{1/2} [\mathcal{M}(P_{h,k}(t-s)) B^* P_{h,k}(t-s) y_{h,k}(s) + u(s)] \right\|^2 \right. \\ &\quad \left. - \left\| \sqrt{S} y_{h,k}(s) \right\|^2 - \|u(s)\|^2 \right\} ds. \end{aligned} \quad (3.2)$$

Thus (3.1) follows from the equality

$$(3.3) \quad \mathbf{E} \left(\sum_{i=1}^{+\infty} \sqrt{\lambda_i} \int_0^t \phi_i(s) d\beta_i^{(1)}(s) + \sum_{j=1}^{+\infty} \sqrt{\nu_j} \int_0^t \psi_j(s) d\beta_j^{(2)}(s) \right) = 0.$$

Note that (3.3) requires a proof since it involves stochastic integrals of functions that are not a priori square integrable. To show (3.3), we introduce the following stopping time:

$$\tau_N(\omega) = \inf \left\{ \tau \in [0, t] \text{ s.t. } \sum_{i=1}^{+\infty} \int_0^\tau \lambda_i \phi_i^2(s, \omega) + \nu_i \psi_i^2(s, \omega) ds \geq N \right\}$$

and we set

$$\begin{aligned} I_N(s) &= \sum_{i=1}^{+\infty} \sqrt{\lambda_i} \int_0^s I_{[0, \tau_N]}(\phi_i(\sigma)) d\beta_i^{(1)}(\sigma) + \sum_{j=1}^{+\infty} \sqrt{\nu_j} \int_0^s I_{[0, \tau_N]}(\psi_j(\sigma)) d\beta_j^{(2)}(\sigma), \\ I(s) &= \sum_{i=1}^{+\infty} \sqrt{\lambda_i} \int_0^s \phi_i(\sigma) d\beta_i^{(1)}(\sigma) + \sum_{j=1}^{+\infty} \sqrt{\nu_j} \int_0^s \psi_j(\sigma) d\beta_j^{(2)}(\sigma), \\ F(s) &= \langle P_{h,k}(t-s)y_{h,k}(s), y_{h,k}(s) \rangle - \langle P_{h,k}(t)x, x \rangle + \int_0^s \left\| \sqrt{S} y_{h,k}(\sigma) \right\|^2 + \|u(\sigma)\|^2 d\sigma \\ &- \int_0^s \left\| \left(I + \sum_{i=1}^{+\infty} \nu_i D_i^* P_{h,k}(t-\sigma) D_i \right)^{1/2} [\mathcal{M}(P_{h,k}(t-\sigma)) B^* P_{h,k}(t-\sigma) y_{h,k}(\sigma) + u(\sigma)] \right\|^2 d\sigma. \end{aligned}$$

Since $y_{h,k}$ has continuous paths, F also has continuous paths, and

$$\mathbf{E} \left(\sup_{s \leq t} |F(s)| \right) < +\infty.$$

By (3.2), $I(s) = F(s)$ almost surely; therefore, choosing the continuous version of I and I_N , we have

$$\mathbf{E}|I(t) - I_N(t)| = \mathbf{E}|I(t) - I(\tau_N)| = \mathbf{E}|F(t) - F(\tau_N)| \rightarrow 0.$$

Hence, since $\mathbf{E} \left(\sum_{i=1}^{+\infty} \int_0^T I_{[0, \tau_N]}(\lambda_i \phi_i^2(\sigma) + \nu_j \psi_j^2(\sigma)) d\sigma \right) < \infty$, we can conclude that

$$\mathbf{E}(I(t)) = \lim_{N \rightarrow +\infty} \mathbf{E}(I_N(t)) = 0,$$

and then (3.3) follows. \square

We can now solve the finite horizon problem in a standard way, as shown below.

THEOREM 3.2. Fix $T > 0$ and $x \in \mathbf{L}^2(\Omega, \mathcal{F}_0, H)$. Then

(i) There exists a unique control $\hat{u} \in \mathbf{M}_{\mathcal{P}}^2(0, T, H)$ such that

$$J_T(x, \hat{u}) = \inf_{u \in \mathbf{M}_{\mathcal{P}}^2(0, T, H)} J_T(x, u).$$

(ii) If \hat{y} is the mild solution of (S.E.) relative to x and \hat{u} , then $\hat{u}(s) = -\mathcal{M}(P(T-s))B^*(T-s)\hat{y}(s)$.

(iii) The optimal cost is given by $J_T(x, \hat{u}) = \mathbf{E}\langle P(T)x, x \rangle$.

Proof. The proof is standard (see [4] and [10]); let us recall it very briefly. Let \hat{y} be the mild solution of

$$(3.4) \quad \begin{aligned} d_s \hat{y}(s) &= (A - BM(P(T-s))B^*P(T-s))\hat{y}(s)ds + C\hat{y}(s)dW_s^{(1)} \\ &\quad - DM(P(T-s))B^*P(T-s)\hat{y}(s)dW_s^{(2)}, \\ \hat{y}(0) &= x, \end{aligned}$$

and let $\hat{u}(s) = -\mathcal{M}(P(T-s))B^*P(T-s)\hat{y}(s)$. From Theorem (3.1) it follows that $J_T(x, \hat{u}) = \mathbf{E}\langle P(T)x, x \rangle$ and that $J_T(x, u) \geq \mathbf{E}\langle P(T)x, x \rangle$ for all $u \in \mathbf{M}_{\mathcal{P}}^2(0, T, H)$.

Conversely, by Theorem (3.1) we deduce that, if u_1 is a control verifying $J_T(x, u_1) = \mathbf{E}\langle P(T)x, x \rangle$ and y_1 is the mild solution of (S.E.) corresponding to x and u_1 , then y_1 is a mild solution of (3.4). Therefore $y_1 = \hat{y}$ and $u_1 = \hat{u}$. \square

4. Synthesis of the optimal control (infinite horizon case).

DEFINITION. We say that (A, B, C, D) is stabilizable relatively to the observations \sqrt{S} (or \sqrt{S} -stabilizable) if, for all $x \in \mathbf{L}^2(\Omega, \mathcal{F}_0, H)$, there exists $u \in \mathbf{M}_{\mathcal{P}}^2(0, T, H)$ such that $J_\infty(x, u) < +\infty$ (such a control is called admissible relatively to x and \sqrt{S}).

Note that this definition is a priori weaker than the one considered in [9], [10], and [5]; Theorem 4.1 and Remark 4.1, appearing later in this section, will show that these two definitions are indeed equivalent.

DEFINITION. We say that $X \in \Sigma^+(H)$ is a mild solution of the algebraic Riccati equation

$$(A.R.E.) \quad A^*X + XA + \sum_{i=1}^{+\infty} \lambda_i C_i^* X C_i - XB \left[I + \sum_{i=1}^{+\infty} \nu_i D_i^* X D_i \right]^{-1} B^*X + S = 0,$$

if it is a mild stationary solution of (R.E.). By \underline{P} we denote, as before, the solution of (R.E.) with initial data $\underline{P}(0) = 0$.

PROPOSITION 4.1. If there exists $\beta \geq 0$ such that $\underline{P}(t) \leq \beta I$, for all $t \geq 0$, then there exists $P_\infty \in \Sigma^+(H)$ such that $\underline{P}(t)x \rightarrow P_\infty x$, for all $x \in H$, and P_∞ is a mild solution of (A.R.E.).

Proof. Under the hypothesis, the existence of the limit P_∞ is a standard consequence of the monotonicity of \underline{P} . We show now that P_∞ is a mild solution of (A.R.E.).

Let \hat{P} be the mild solution of (R.E.) with initial data $\hat{P}(0) = P_\infty$ and, for all $N \in \mathbb{N}$, let P_N be the mild solution of (R.E.) with initial data $P_N(0) = \underline{P}(N)$. Proceeding as in the proof of Theorem 2.1, point (iii), we can show the continuous dependence of the mild solutions of (R.E.) on initial data. This implies, in our particular case, that $P_N x \rightarrow \hat{P}x$

for all $x \in H$ uniformly in the bounded subsets of $[0, +\infty[$. Also, for all $t \geq 0$, we have $P_N(t) = P(t + N)$ and then, letting $N \rightarrow +\infty$, we get $\hat{P}(t) = P_\infty$ for all $t \geq 0$, and this proves that P_∞ is a mild solution of $(\mathcal{A.R.E.})$. \square

Note that P_∞ is the minimal positive mild solution of $(\mathcal{A.R.E.})$ (see Theorem (2.2)). The next theorem is the main result of this section.

THEOREM 4.1. *The three following assertions are equivalent:*

- (i) (A, B, C, D) is stabilizable relatively to \sqrt{S} ,
- (ii) There exists $\beta \geq 0$ such that, for all $t \geq 0$, $\underline{P}(t) \leq \beta I$,
- (iii) There exists a mild solution of $(\mathcal{A.R.E.})$.

Proof. The equivalence of (ii) and (iii) is an obvious consequence of Proposition 4.1 and of Theorem 2.2. Let us prove that (i) \Rightarrow (ii).

Denote by x_h ($h \in H$) the random function identically equal to the constant h and let u_h be an admissible control relatively to x_h and \sqrt{S} . Then by Theorem 3.1, it follows that $\langle \underline{P}(t)h, h \rangle \leq J_\infty(x_h, u_h)$ for all $t \geq 0$. So we have proved that, for all $h \in H$, $\langle \underline{P}(t)h, h \rangle$ is bounded; from this it follows by using the Baire category argument, that there exists $\beta \geq 0$ such that $\underline{P}(t) \leq \beta I$, for all $t \geq 0$.

Now let us see that (ii) \Rightarrow (i).

Fix $x \in L^2(\Omega, \mathcal{F}_0, \mathbf{P})$ and let P_∞ be, as before, the minimal mild solution of $(\mathcal{A.R.E.})$. Moreover let, for all $N \in \mathbb{N}$, y_N be the mild solution of

$$\begin{aligned} d_s y_N(s) &= (A - B\mathcal{M}(\underline{P}(N-s))B^* \underline{P}(N-s))y_N(s)ds + Cy_N(s)dW_s^{(1)} \\ &\quad - D\mathcal{M}(\underline{P}(N-s))B^* \underline{P}(N-s)y_N(s)dW_s^{(2)}, \\ y_N(0) &= x; \end{aligned}$$

let \hat{y} be the mild solution of

$$\begin{aligned} d_s \hat{y}(s) &= (A - B\mathcal{M}(P_\infty)B^* P_\infty)\hat{y}(s)ds + C\hat{y}(s)dW_s^{(1)} \\ &\quad - D\mathcal{M}(P_\infty)B^* P_\infty \hat{y}(s)dW_s^{(2)}, \\ \hat{y}(0) &= x; \end{aligned} \tag{4.1}$$

and let

$$\begin{aligned} u_N(s) &= -\mathcal{M}(\underline{P}(N-s))B^* \underline{P}(N-s)y_N(s); \\ \hat{u}(s) &= -\mathcal{M}(P_\infty)B^* P_\infty \hat{y}(s). \end{aligned} \tag{4.2}$$

As we have done in Remark 1.3, we can prove that $y_N \rightarrow \hat{y}$ in $\mathbf{M}_{\mathcal{P}}^2(0, T, V)$ (for all $T \geq 0$) and, therefore, that $u_N \rightarrow \hat{u}$ in $\mathbf{M}_{\mathcal{P}}^2(0, T, H)$.

Fix T ; if $N \geq T$, Theorem 3.1 yields

$$\mathbf{E}\langle P_\infty x, x \rangle \geq \mathbf{E}\langle \underline{P}(N)x, x \rangle \geq \mathbf{E} \int_0^T \left\| \sqrt{S}y_N(s) \right\|^2 + \|u_N(s)\|^2 ds. \tag{4.3}$$

So, letting $N \rightarrow +\infty$ in (4.3), we get

$$\mathbf{E}\langle P_\infty x, x \rangle \geq \mathbf{E} \int_0^T \left\| \sqrt{S}\hat{y}(s) \right\|^2 + \|\hat{u}(s)\|^2 ds \quad \forall T \geq 0. \tag{4.4}$$

Then \hat{u} is the admissible control we were seeking. \square

Let us note that, letting $T \rightarrow +\infty$ in (4.4), we obtain

$$(4.5) \quad \mathbf{E}(P_\infty x, x) \geq J_\infty(x, \hat{u}).$$

Remark 4.1. If \hat{y} is the mild solution of (4.1) corresponding to $x \in L^2(\Omega, \mathcal{E}, \mathbf{P})$, then

$$\mathbf{E} \int_0^T \left\| \sqrt{S} \hat{y}(s) \right\|^2 ds < +\infty.$$

Thus, if (A, B, C, D) is stabilizable, then it is stabilizable “by a feedback” (see [5] and [10]).

Remark 4.2. To prove that (A, B, C, D) is stabilizable relatively to \sqrt{S} , it is enough to show that, for all constant initial data $x_0(\omega) = h$ ($h \in H$ and $\omega \in \Omega$), there exists an admissible control relatively to x_0 and \sqrt{S} .

Proof. The proof follows from the proof of (i) \Rightarrow (ii) in Theorem 4.1. \square

Remark 4.3. Suppose that there exists $B^{-1} \in \mathcal{L}(H)$; then there exists $d_0 \geq 0$ (d_0 depends on A, B , and C) such that

$$\sum_{i=1}^{\infty} \nu_i \|D_i\|^2 < d_0 \Rightarrow (A, B, C, D) \text{ is stabilizable relatively to the identity.}$$

Proof. First, let $D_i = 0$ (for all $i \in \mathbb{N}$) and set $\tilde{A} = A - m$.

Now if m is large enough and if, for all $x \in H$, ξ_x is the mild solution of

$$\begin{aligned} d\xi_x &= \tilde{A}\xi_x dt + \sum_{i=1}^{\infty} \lambda_i^{1/2} C_i \xi_x d\beta_i(t), \\ \xi_x(0) &= x, \end{aligned}$$

then, applying the techniques we used to prove Theorem 1.1, we get

$$\mathbf{E} \int_0^{+\infty} \|\xi_x(t)\|^2 dt < \infty.$$

Therefore if $u = -mB^{-1}\xi_x$, then u is admissible relatively to x and the identity. We have therefore shown that $(A, B, C, 0)$ is I -stabilizable. Therefore there exists $X \in \Sigma^+(H)$ mild solution of

$$A^*X + XA + \sum_{i=1}^{+\infty} \lambda_i C_i^* X C_i - XBB^*X + I = 0.$$

The previous equation can be rewritten as

$$A^*X + XA + \sum_{i=1}^{+\infty} \lambda_i C_i^* X C_i - XBM(X)B^*X + \tilde{S} = 0,$$

where

$$\tilde{S} = -XBM(X)B^*X + XBB^*X + I.$$

Thus if $\tilde{S} \geq 0$, then (A, B, C, D) is $\sqrt{\tilde{S}}$ -stabilizable and, in particular, if $\tilde{S} \geq \alpha I$ (for some $\alpha > 0$), then (A, B, C, C) is I -stabilizable.

The proof is then completed considering the continuity of \mathcal{M} . \square

Now we solve the infinite horizon problem as in [4] and [10].

THEOREM 4.2. Suppose that (A, B, C, D) is stabilizable relatively to \sqrt{S} and fix any $x \in L^2(\Omega, \mathcal{F}_0, H)$. Then

(i) *There exists a unique control $\hat{u} \in \mathbf{M}_{\mathcal{P}}^2(0, +\infty, H)$ such that*

$$J_{\infty}(x, \hat{u}) = \inf_{u \in \mathbf{M}_{\mathcal{P}}^2(0, +\infty, H)} J_{\infty}(x, u),$$

(ii) *If \hat{y} is the mild solution of (S.E.) corresponding to x and \hat{u} , then*

$$\hat{u}(s) = -\mathcal{M}(P_{\infty})B^*P_{\infty}\hat{y}(s),$$

(iii) *The optimal cost is given by $J_{\infty}(x, \hat{u}) = \mathbf{E}\langle P_{\infty}x, x \rangle$.*

Proof. Theorem 3.1 yields (letting $t \rightarrow +\infty$)

$$J_{\infty}(x, u) \geq \mathbf{E}\langle P_{\infty}x, x \rangle \quad \forall u \in \mathbf{M}_{\mathcal{P}}^2(0, +\infty, H).$$

Therefore if \hat{u} and \hat{y} are defined as in (4.1) and (4.2), from (4.5) we deduce that

$$J_{\infty}(x, \hat{u}) = \mathbf{E}\langle P_{\infty}x, x \rangle = \inf_{u \in \mathbf{M}_{\mathcal{P}}^2(0, +\infty, H)} J_{\infty}(x, u).$$

We now show that the optimal control is unique.

Let $u_1 \in \mathbf{M}_{\mathcal{P}}^2(0, +\infty, H)$ such that $J_{\infty}(x, u_1) = \mathbf{E}\langle P_{\infty}x, x \rangle$ and let y_1 be the corresponding mild solution of (S.E.). Theorem 3.1 yields

$$\int_0^N \|\mathcal{M}(\underline{P}(N-s))B^*\underline{P}(N-s)y_1(s) + u_1(s)\|^2 ds \leq J_{\infty}(x, u_1) - \mathbf{E}\langle \underline{P}(N)x, x \rangle.$$

Thus if $T \leq N$,

$$\int_0^T \|\mathcal{M}(\underline{P}(N-s))B^*\underline{P}(N-s)y_1(s) + u_1(s)\|^2 ds \leq \mathbf{E}\langle (P_{\infty} - \underline{P}(N))x, x \rangle,$$

and then, letting $N \rightarrow +\infty$, we get

$$u_1 = -\mathcal{M}(P_{\infty})B^*P_{\infty}y_1 \quad \text{a.e. in } \Omega \times [0, T],$$

and, finally, $y_1 = \hat{y}$; $u_1 = \hat{u}$. \square

5. Attractivity and maximality properties of the stabilizing solutions of (A.R.E.).

From now on, we suppose that (A, B, C, D) is stabilizable with respect to \sqrt{S} .

DEFINITION. We say that a solution $X \in \Sigma^+(H)$ of (A.R.E.) stabilizes (A, B, C, D) relatively to I if the triple $(A - B\mathcal{M}(X)B^*X, C, -D\mathcal{M}(X)B^*X)$ is stable; that is, if, for all $x \in \mathbf{L}^2(\Omega, \mathcal{F}_0, \mathbf{P})$, we have $\mathbf{E} \int_0^{+\infty} \|\xi_x(s)\|^2 ds < +\infty$, where ξ_x is the mild solution of

$$(5.1) \quad \begin{aligned} d\xi_x(s) &= (A - B\mathcal{M}(X)B^*X)\xi_x(s)ds + C\xi_x dW_s^{(1)} - D\mathcal{M}(X)B^*X\xi_x dW_s^{(2)}, \\ \xi_x(0) &= x. \end{aligned}$$

PROPOSITION 5.1. *Let X and Y be two positive mild solutions of (A.R.E.) and suppose that X stabilizes (A, B, C, D) relatively to I ; then $X \geq Y$.*

Proof. Let $x \in \mathbf{L}^2(\Omega, \mathcal{F}_0, H)$ and let ξ_x be the mild solution of (5.1). From Theorem 3.1, setting $Z = X - Y$, it follows that

$$\begin{aligned} \mathbf{E}\langle Z\xi_x(t), \xi_x(t) \rangle - \mathbf{E}\langle Zx, x \rangle \\ = -\mathbf{E} \int_0^t \left\| \left(I + \sum_{i=1}^{+\infty} \nu_i D_i^* Y D_i \right)^{1/2} [\mathcal{M}(Y)B^*Y - \mathcal{M}(X)B^*X]\xi_x(s) \right\|^2 ds. \end{aligned}$$

Then, in particular, choosing a constant initial data $x_h = h \in H$, we have

$$(5.2) \quad \langle Zh, h \rangle \geq \mathbf{E} \langle Z\xi_x(t), \xi_x(t) \rangle.$$

Since X stabilizes (A, B, C, D) relatively to I , we know that $\mathbf{E} \int_0^{+\infty} \|\xi_x(s)\|^2 ds < +\infty$. Moreover, being ξ_x continuous in the $\mathbf{L}^2(\Omega, \mathcal{E}, \mathbf{P}, H)$ norm, we can find a sequence $\{t_n : n \in \mathbb{N}\}$ such that $\lim_{n \rightarrow +\infty} \mathbf{E} \|\xi_x(t_n)\|^2 = 0$. Substituting t_n in (5.2) and letting $n \rightarrow +\infty$, we get $\langle Zh, h \rangle \geq 0$. \square

A trivial consequence of Proposition 5.1 is the following corollary.

COROLLARY 5.1. *If the minimal mild solution P_∞ of the above $(A, \mathcal{R}, \mathcal{E})$ stabilizes (A, B, C, D) relatively to I , then it is the unique mild solution of the $(A, \mathcal{R}, \mathcal{E})$.*

The following proposition is proved, with the same technique, in [5].

PROPOSITION 5.2. *Let $X \in \Sigma^+(H)$ be a mild solution of $(A, \mathcal{R}, \mathcal{E})$ and suppose that X stabilizes (A, B, C, D) relatively to I . Then there exists $M \geq 1$ and $\omega_0 > 0$ such that for all mild solutions P of $(\mathcal{R}, \mathcal{E})$, verifying $P(0) = P_0 \geq X$, the following estimate holds:*

$$\|P(t) - X\| \leq M e^{-\omega_0 t} \|P_0 - X\| \quad \forall t \geq 0.$$

Proof. Again, let ξ_x be the mild solution of (5.1) (for all $x \in \mathbf{L}^2(\Omega, \mathcal{F}_0, H)$) and, again, take for all $\omega \in \Omega$, $x(\omega) = h \in H$. Setting $T(t) = P(t) - X$ from Theorem 3.1, it follows that, for all $t \geq 0$,

$$\begin{aligned} & \mathbf{E} \langle T(0)\xi_x(t), \xi_x(t) \rangle - \mathbf{E} \langle T(t)h, h \rangle \\ &= \mathbf{E} \int_0^t \left\| \left(I + \sum_{i=1}^{+\infty} \nu_i D_i^* P(t-s) D_i \right)^{1/2} [\mathcal{M}(P(t-s)) B^* P(t-s) - \mathcal{M}(X) B^* X] \xi_x(s) \right\|^2 ds. \end{aligned}$$

Therefore, being $T(t) \geq 0$ (see Theorem (2.2)), we deduce that

$$(5.3) \quad 0 \leq \langle T(t)h, h \rangle \leq \mathbf{E} \langle T(0)\xi_x(t), \xi_x(t) \rangle.$$

Since $\mathbf{E} \int_0^{+\infty} \|\xi_x(s)\|^2 ds < +\infty$ for all $x \in \mathbf{L}^2(\Omega, \mathcal{F}_0, H)$ by the stochastic version of the Datko theorem (see [5, Prop. 2.2] and [11 Thm. 2.1]), it follows that there exists $M \geq 1$ and $\omega_0 > 0$ such that, for all $x \in \mathbf{L}^2(\Omega, \mathcal{F}_0, H)$, $t \geq 0$,

$$(5.4) \quad \mathbf{E} \|\xi_x(t)\|^2 \leq M e^{-\omega_0 t} \mathbf{E} \|x\|^2.$$

The claim is now a trivial consequence of (5.3) and (5.4). \square

COROLLARY 5.2. *If the minimal mild solution P_∞ of the $(A, \mathcal{R}, \mathcal{E})$ stabilizes (A, B, C, D) relatively to I , then it is globally attractive among all the (positive) mild solutions of the corresponding $(\mathcal{R}, \mathcal{E})$.*

Proof. Let P be any mild solution of $(\mathcal{R}, \mathcal{E})$ and let $\beta \in \mathbb{R}$ such that $P(0) \leq \beta I$ and $P_\infty \leq \beta I$. Denote by Q the mild solution of $(\mathcal{R}, \mathcal{E})$ with $Q(0) = \beta I$. From Theorem 2.2, it follows that, for all $t \geq 0$,

$$(5.5) \quad 0 \leq P(t) - \underline{P}(t) \leq Q(t) - \underline{P}(t).$$

Now since $\underline{P}(t)h \rightarrow P_\infty h$ and $Q(t)h \rightarrow P_\infty h$, for all $h \in H$, (see Prop. 5.2) the claim follows from (5.5). \square

COROLLARY 5.3. *If $S \geq \beta I$ for some $\beta < 0$, then the $(A, \mathcal{R}, \mathcal{E})$ has a unique mild solution that is globally attractive among all the (positive) mild solutions of the corresponding $(\mathcal{R}, \mathcal{E})$ (remember that we are still assuming that (A, B, C, D) is stabilizable relatively to \sqrt{S}).*

Proof. It is an easy consequence of Corollaries 5.1 and 5.2 because, under this hypothesis about S , Theorem 4.1 and Remark 4.1 imply that the minimal mild solution P_∞ of the $(A, \mathcal{R}, \mathcal{E})$ stabilizes (A, B, C, D) relatively to I . \square

6. Maximal solution of the algebraic Riccati equation. From now on, we suppose that (A, B, C, D) is stabilizable relatively to the identity. Fix $\epsilon > 0$; let $S_\epsilon = S + \epsilon I$ and let X_ϵ be the mild solution of the following algebraic Riccati equation:

$$(6.1) \quad A^* X_\epsilon + X_\epsilon A + \sum_{i=1}^{+\infty} \lambda_i C_i^* X_\epsilon C_i + S_\epsilon - X_\epsilon B \left[I + \sum_{i=1}^{+\infty} \nu_i D_i^* X_\epsilon D_i \right]^{-1} B^* X_\epsilon = 0.$$

Corollary 5.3 implies that (6.1) has a unique mild solution X_ϵ , which is globally attractive among the mild solutions of the corresponding Riccati equation.

By Theorem 2.2, it follows that if $0 \leq \epsilon_1 \leq \epsilon_2$, then $0 \leq X_{\epsilon_1} \leq X_{\epsilon_2}$; therefore there exists $X \in \Sigma^+(H)$ such that, for all $h \in H$, $X_\epsilon h \rightarrow Xh$ (if $\epsilon \rightarrow 0$).

PROPOSITION 6.1. *X is the maximal mild solution of the following algebraic Riccati equation:*

$$(6.2) \quad A^* X + X A + \sum_{i=1}^{+\infty} \lambda_i C_i^* X C_i + S - X B \left[I + \sum_{i=1}^{+\infty} \nu_i D_i^* X D_i \right]^{-1} B^* X = 0.$$

Proof. To see that X is a mild stationary solution of $(\mathcal{R}, \mathcal{E})$, it is enough to let $\epsilon \rightarrow 0$ in the following relation:

$$\begin{aligned} X_\epsilon x_0 &= e^{tA^*} X_\epsilon e^{tA} x_0 + \sum_{i=1}^{+\infty} \lambda_i \int_0^t \left(C_i e^{(t-s)A} \right)^* X_\epsilon \left(C_i e^{(t-s)A} \right) x_0 ds \\ &\quad + \int_0^t e^{(t-s)A^*} S_\epsilon e^{(t-s)A} x_0 ds \\ &\quad - \int_0^t e^{(t-s)A^*} X_\epsilon B \left[I + \sum_{i=1}^{+\infty} \nu_i D_i^* X_\epsilon D_i \right]^{-1} B^* X_\epsilon e^{(t-s)A} x_0 ds, \end{aligned}$$

which, by the definition of X_ϵ , holds for all $x_0 \in H$, $t > 0$, and $\epsilon > 0$.

We claim that X is maximal. Let Q be another mild solution of (6.2) and let P_ϵ be the solution of

$$P'_\epsilon = A^* P_\epsilon + P_\epsilon A + \sum_{i=1}^{+\infty} \lambda_i C_i^* P_\epsilon C_i + S_\epsilon - P_\epsilon B \left[I + \sum_{i=1}^{+\infty} \nu_i D_i^* P_\epsilon D_i \right]^{-1} B^* P_\epsilon,$$

$$P_\epsilon(0) = Q.$$

Then Theorem (2.2) yields $P_\epsilon(t) \geq Q$, for all $t \geq 0$. Moreover, from the global attractivity of X_ϵ , we deduce that $P_\epsilon(t)h \rightarrow X_\epsilon h$ as $t \rightarrow +\infty$ (for any fixed $\epsilon > 0$ and any $h \in H$). We complete the proof by considering that, by definition, $X_\epsilon h \rightarrow Xh$ as $\epsilon \rightarrow 0$ (for all $h \in H$). \square

We now want to show that the maximal solution is related to a special kind of control problem (see [2] for the deterministic case). Let, for all $x \in \mathbf{L}^2(\Omega, \mathcal{F}_0, \mathbf{P})$, $U_{ad}(x)$ be the set of all controls $u \in \mathbf{M}_{\mathbf{P}}^2(0, +\infty, H)$ such that the mild solution of (S, \mathcal{E}) corresponding to x and u belongs to $\mathbf{M}_{\mathbf{P}}^2(0, +\infty, H)$. Note that, for all $x \in \mathbf{L}^2(\Omega, \mathcal{F}_0, \mathbf{P})$, $U_{ad}(x)$ is nonempty.

PROPOSITION 6.2. *For every $x \in \mathbf{L}^2(\Omega, \mathcal{F}_0, \mathbf{P}, H)$, we have*

$$(6.3) \quad \mathbf{E}\langle Xx, x \rangle = \inf_{u \in U_{ad}(x)} J_{\infty}(x, u).$$

Proof. Let $x \in \mathbf{L}^2(\Omega, \mathcal{F}_0, \mathbf{P}, H)$ and $u \in U_{ad}(x)$. Then, for all $t \geq 0$, Theorem 3.1 yields

$$(6.4) \quad \mathbf{E}\langle Xx, x \rangle \leq \mathbf{E} \int_0^t \left(\left\| \sqrt{S}y(\sigma) \right\|^2 + \|u(\sigma)\|^2 \right) d\sigma + \mathbf{E}\langle Xy(t), y(t) \rangle,$$

where y is the mild solution of (S, \mathcal{E}) corresponding to x and u . Since $\mathbf{E}(\|y\|^2) \in \mathcal{C}(0, +\infty, \mathbf{R}) \cap \mathbf{L}^2(0, +\infty, \mathbf{R})$ we can find a sequence $\{t_n\}$ such that $\lim_{n \rightarrow +\infty} t_n = +\infty$ and $\lim_{n \rightarrow +\infty} \mathbf{E}(\|y(t_n)\|^2) = 0$. Setting $t = t_n$ in (6.4) and letting $n \rightarrow +\infty$, we get

$$(6.5) \quad \mathbf{E}\langle Xx, x \rangle \leq J_{\infty}(x, u).$$

Now if y_{ϵ} denotes the mild solution of

$$(6.6) \quad \begin{aligned} dy_{\epsilon} &= (A - B\mathcal{M}(X_{\epsilon})B^*X_{\epsilon})y_{\epsilon}ds + Cy_{\epsilon}dW_s^{(1)} - D\mathcal{M}(X_{\epsilon})B^*X_{\epsilon}y_{\epsilon}dW_s^{(2)}, \\ y_{\epsilon}(0) &= x, \end{aligned}$$

and we set $u_{\epsilon} = -\mathcal{M}(X_{\epsilon})B^*X_{\epsilon}y_{\epsilon}$, we have from Theorem 4.1 and Remark 4.1 that u_{ϵ} belongs to $U_{ad}(x)$ and that

$$(6.7) \quad \mathbf{E}\langle X_{\epsilon}x, x \rangle \geq J_{\infty}(x, u_{\epsilon}) + \epsilon \int_0^{+\infty} \mathbf{E}\|y_{\epsilon}(\sigma)\|^2 d\sigma.$$

From (6.5) and (6.7), we deduce that

$$(6.8) \quad \epsilon \int_0^{+\infty} \mathbf{E}\|y_{\epsilon}(\sigma)\|^2 d\sigma \leq \mathbf{E}\langle (X_{\epsilon} - X)x, x \rangle.$$

From (6.8) it follows, by the dominated convergence theorem, that

$$\epsilon \int_0^{+\infty} \mathbf{E}\|y_{\epsilon}(\sigma)\|^2 d\sigma \rightarrow 0.$$

The proof is complete if we substitute this in (6.7) and let $\epsilon \rightarrow 0$. \square

PROPOSITION 6.3. *Let X be the maximal mild solution of (6.2). The following statements are equivalent:*

- (i) $\mathbf{E}\langle Xx, x \rangle = \min_{u \in U_{ad}(x)} J_{\infty}(x, u)$ for all $x \in \mathbf{L}^2(\Omega, \mathcal{F}_0, \mathbf{P}, H)$,
- (ii) X stabilizes (A, B, C, D) relatively to I .

Proof. Fix $x \in \mathbf{L}^2(\Omega, \mathcal{F}_0, \mathbf{P})$ and let $u \in U_{ad}(x)$ such that $\mathbf{E}\langle Xx, x \rangle = J_\infty(x, u)$; let y be the mild solution of $(S.E.)$ corresponding to x and u . Then from Theorem 3.1 it follows that

$$(6.9) \quad \begin{aligned} & \mathbf{E}\langle Xx, x \rangle + \mathbf{E} \int_0^t \left\| \left(I + \sum_{i=1}^{+\infty} \nu_i D_i^* X D_i \right)^{1/2} [\mathcal{M}(X) B^* X y(s) + u(s)] \right\|^2 ds \\ &= \mathbf{E}\langle Xy(t), y(t) \rangle + \mathbf{E} \int_0^t \left(\|\sqrt{S}y(s)\|^2 + \|u(s)\|^2 \right) ds. \end{aligned}$$

Letting $t \rightarrow +\infty$ (this can be justified as in Proposition 6.2), we get

$$\mathbf{E} \int_0^{+\infty} \left\| \left(I + \sum_{i=1}^{+\infty} \nu_i D_i^* X D_i \right)^{1/2} [\mathcal{M}(X) B^* X y(s) + u(s)] \right\|^2 ds = 0.$$

Hence $u = -\mathcal{M}(X) B^* X y$ and y is a mild solution of (5.1); moreover, we already know that y belongs to $\mathbf{M}_{\mathcal{P}}^2(0, +\infty, H)$. Since this argument holds for every $x \in \mathbf{L}^2(\Omega, \mathcal{F}_0, \mathbf{P})$, we have proved that X stabilizes (A, B, C, D) relatively to I .

Conversely, if X stabilizes (A, B, C, D) relatively to I , let \hat{y} be the mild solution of

$$\begin{aligned} d\hat{y} &= (A - B\mathcal{M}(X)B^*X)\hat{y}ds + C\hat{y}dW_s^{(1)} - D\mathcal{M}(X)B^*X\hat{y}dW_s^{(2)}, \\ \hat{y}(0) &= x, \end{aligned}$$

and let $\hat{u} = -\mathcal{M}(X)B^*X\hat{y}$. Then $\hat{u} \in U_{ad}(x)$, and if we substitute \hat{u} and \hat{y} in (6.9) and we let $t \rightarrow +\infty$, we get (with the usual argument) that $J_\infty(x, \hat{u}) = \mathbf{E}\langle Xx, x \rangle$, and this is exactly what we wanted to prove. \square

7. Example. We specify here the assumptions of the example stated in the Introduction. We consider the following parabolic equation:

$$\begin{aligned} dy &= \left[\sum_{i,j=1}^d \frac{\partial}{\partial x_i} \left(a_{i,j}(x) \frac{\partial y(x)}{\partial x_j} \right) + a(x)y(x) + b(x)u(t, x) \right] dt \\ &\quad + \left[\sum_{i=1}^d c_i(x) \frac{\partial y(x)}{\partial x_i} + c(x)y(x) \right] dw_t^{(1)} + d(x)u(t, x)dw_t^{(2)}, \quad x \in \mathcal{O}; \\ \frac{\partial y(t, x)}{\partial \nu_{\mathbf{a}}} &= 0, \quad x \in \partial\mathcal{O}; \\ y(0, x) &= y_0(x), \quad x \in \mathcal{O}, \end{aligned}$$

where \mathcal{O} is a bounded domain in \mathbb{R}^d with smooth boundary; $w^{(1)}$ and $w^{(2)}$ are two independent standard (one-dimensional) Brownian motions, and $\partial y(t, x)/\partial \nu_{\mathbf{a}}$ denotes the conormal derivative

$$\frac{\partial y(x)}{\partial \nu_{\mathbf{a}}} = \sum_{i,j=1}^d a_{i,j}(x) \frac{\partial y(x)}{\partial x_i} \nu_j,$$

ν being the outward normal to $\partial\mathcal{O}$. We suppose that $a_{i,j} \in \mathcal{C}^1(\bar{\mathcal{O}})$ (for all $i, j \in \mathbb{N}$), a, b, c, d, c_i , all belong to $\mathbf{L}^\infty(\mathcal{O})$. We will also assume that there exists a constant $\ell > 0$,

verifying

$$\sum_{i,j=1}^d (a_{i,j}(x)\xi_i\xi_j - \frac{1}{2}c_i(x)c_j(x)\xi_i\xi_j) \geq \ell\|\xi\|^2 \quad \forall \xi \in \mathbb{R}^d; x \in \mathcal{O}.$$

Now if we take $V = \mathbf{H}^1(\mathcal{O})$, $H = \mathbf{L}^2(\mathcal{O})$ and define on V the bilinear form

$$\mathbf{a}(u, v) = \int_{\mathcal{O}} \sum_{i,j=1}^d a_{i,j}(x) \frac{\partial u}{\partial x_i}(x) \frac{\partial v}{\partial x_j}(x) dx - \int_{\mathcal{O}} a(x) u(x) v(x) dx,$$

then $-\mathbf{a}$ is regularly dissipative. Moreover, if A is the linear operator relative to $-\mathbf{a}$, V , and H (see §1), then we have exactly (see [1])

$$\begin{aligned} \mathcal{D}(A) &= \left\{ u \in \mathbf{H}^2(\mathcal{O}) : \frac{\partial u(x)}{\partial \nu_{\mathbf{a}}} = 0 \text{ on } \partial\mathcal{O} \right\}, \\ (Au)(x) &= \sum_{i,j=1}^d \frac{\partial}{\partial x_i} \left(a_{i,j}(x) \frac{\partial u(x)}{\partial x_j} \right) + a(x) u(x). \end{aligned}$$

Moreover, if $Cv = \sum_{i=1}^{+\infty} c_i(\partial v / \partial x_i) + cv$, $Du = du$, $Bv = bv$, then (Hyp. 1) are verified. Note that

$$\mathcal{D}(A^*) = \left\{ u \in \mathbf{H}^2(\mathcal{O}) : \frac{\partial u(x)}{\partial \nu_{\mathbf{a}^*}} = 0 \text{ on } \partial\mathcal{O} \right\},$$

where

$$\frac{\partial u(x)}{\partial \nu_{\mathbf{a}^*}} = \sum_{i,j=1}^d a_{i,j}(x) \frac{\partial y(x)}{\partial x_j} \nu_i.$$

Thus, in general, $\mathcal{D}(A) \neq \mathcal{D}(A^*)$. So the results proved in [3] and [5] cannot be applied to this equation.

If we consider a finite-horizon cost functional such as

$$J_T(x, u) = \mathbf{E} \int_0^T \int_{\mathcal{O}} s(x) |y(x, t)|^2 + |u(x, t)|^2 dx dt + \mathbf{E} \int_{\mathcal{O}} p(x) |y(x, T)|^2 dx,$$

where $s, p \in \mathbf{L}^\infty(\mathcal{O})$ and $s, p \geq 0$ almost everywhere, then all our hypotheses are fulfilled, and we can conclude that there exists a unique optimal control and that it verifies the feedback law stated in Theorem 3.2. Let us now consider an infinite-horizon cost functional such as

$$J_\infty(x, u) = \mathbf{E} \int_0^\infty \int_{\mathcal{O}} s(x) |y(x, t)|^2 + |u(x, t)|^2 dx dt.$$

Note that if we suppose that $b^{-1} \in \mathbf{L}^\infty(\mathcal{O})$ and that $d = 0$, then we know by Remark 4.3 that the state equation we are considering is I -stabilizable. Therefore we can conclude that there exists a unique optimal control and that it verifies the feedback law stated in Theorem 4.2. Moreover, the $(\mathcal{A}, \mathcal{R}, \mathcal{E})$ corresponding to our problem has a minimal and a maximal positive mild solution, and, if $s(x) \geq s_0$ for some $s_0 > 0$, they coincide.

REFERENCES

- [1] S. AGMON, *Lectures on Elliptic Boundary Value Problems*, van Nostrand, Princeton, NJ, 1965.
- [2] A. BENSOUSSAN, *Observateurs et stabilité*, Colloque CNES, 1987, Paris.
- [3] G. DA PRATO, *Some results on linear stochastic evolution equations in Hilbert spaces by the semi-groups method*, Stochastic Anal. Appl., 1 (1983), pp. 353–375.
- [4] ———, *Direct Solution of a Riccati equation arising in stochastic control theory*, Appl. Math. Optim., 11 (1984), pp. 191–208.
- [5] G. DA PRATO AND A. ICHIKAWA, *Stability and quadratic control for linear stochastic equations with unbounded coefficient*, Boll. Un. Mat. Ital. B (6), 4 (1985), pp. 987–1001.
- [6] ———, *Riccati equations with unbounded coefficients*, Ann. Mat. Pura Appl. (4), 115 (1985), pp. 209–211.
- [7] F. FLANDOLI, *Riccati equation arising in stochastic control theory*, Boll. Un. Mat. Ital.; Anal. Funz. Appl., 1 (1982), pp. 377–393.
- [8] U. G. HAUSMANN, *Optimal stationary control with state and control dependent noise*, SIAM J. Control, 9 (1971), pp. 184–198.
- [9] A. ICHIKAWA, *Optimal control of a linear stochastic evolution equation with state and control dependent noise*, in Proc. IMA Conference “Recent Theoretical Developments in Control,” Leicester, England, Academic Press, 1976.
- [10] ———, *Dynamic programming approach to stochastic evolution equation*, SIAM J. Control Optim., 17 (1979), pp. 152–173.
- [11] ———, *Equivalence of L_p stability and exponential stability for a class of nonlinear semigroups*, Nonlinear Anal., 8 (1984), pp. 805–815.
- [12] N. V. KRYLOV, *Controlled Diffusion Processes*, Springer-Verlag, New York, Heidelberg, Berlin, 1980.
- [13] M. METIVIER, *Semimartingales. A Course on Stochastic Processes*, de Gruyter, Berlin, New York, 1982.
- [14] A. J. PRITCHARD AND J. ZABCZYK, *Stability and stabilizability of infinite-dimensional systems*, SIAM Rev., 23 (1981), pp. 25–52.
- [15] H. TANABE, *Equations of Evolution*, Pitman, London, 1979.
- [16] W. M. WONHAM, *On a matrix Riccati equation of stochastic control*, SIAM J. Control, 6 (1968), pp. 681–697.
- [17] J. ZABCZYK, *Remarks on the algebraic Riccati equation in Hilbert space*, Appl. Math. Optim., 2 (1976), pp. 251–258.

REGULARIZED MAXIMUM LIKELIHOOD ESTIMATE FOR AN INFINITE-DIMENSIONAL PARAMETER IN STOCHASTIC PARABOLIC SYSTEMS*

SHIN ICHI AIHARA†

Abstract. The purpose of this paper is to study the identification problem of an infinite-dimensional parameter, more precisely a spatially varying parameter, in stochastic diffusion equations. In a previous study [S. I. Aihara and Y. Sunahara, *SIAM J. Control Optim.*, 26 (1988), pp. 1062–1075], some explicit conditions for the consistency property of the maximum likelihood estimate (MLE) is explored. Here, an algorithm for generating the MLE is developed with the aid of the regularization technique proposed by [C. Kravaris and J. H. Seinfeld, *SIAM J. Control Optim.*, 23 (1985), pp. 217–241]. After the consistency property of the MLE by a regularization is proved, necessary conditions for the regularized MLE (RMLE) are derived. Proposed is an iterative algorithm for computing one of the solutions of the necessary conditions derived. The convergence property of the sequence generated by the proposed algorithm is also shown. Finally, numerical examples are presented.

Key words. maximum likelihood estimate (MLE), consistent estimate, regularization, stochastic parabolic systems

AMS(MOS) subject classifications. 93C29, 93E12

1. Introduction. This paper deals with the identification of an infinite-dimensional parameter in stochastic parabolic systems. In a previous study [3], sufficient conditions ensuring the consistency of the MLE for a spatially varying parameter are presented. In this paper, our final goal is to construct an algorithm to find one of the solutions that satisfies the necessary conditions for the MLE. In [3] the consistency property of the MLE for the infinite-dimensional parameter in stochastic diffusion equations was derived with the aid of the pioneering work of Bagchi and Borkar [4]. However, the derivation of necessary conditions for the MLE is not straightforward except in the finitely additive white-noise observation case [2]. Before describing the identification problem of stochastic systems, we consider deterministic systems. For the deterministic case, identification problems have been reformulated by converting them into questions of optimal control, regarding unknown parameters as control signals. Necessary conditions for minimizing a certain error criterion are derived, in terms of the so-called “adjoint equation” (see [7], [11]). The generating algorithm for the parameter estimate was derived from these necessary conditions, as used in [7]. For the stochastic case, the adjoint equation should be defined as a backward stochastic equation, in some sense. Noting that stochastic parabolic equations cannot be defined in the time-reversed sense, the formulation of these equations was an open problem. Recently, Benes and Karatzas [5] succeeded in formulating the adjoint equation for a stochastic parabolic system, namely Zakai’s equation, by using the so-called “pathwise method” proposed by Rosovsky (see [13, p. 327]) and the related transformation. In this paper, we derive the necessary conditions for the MLE and construct an algorithm to find a solution of the proposed necessary conditions by transforming the filter equation into the pathwise form. The usual approach for deriving the algorithm to identify unknown

* Received by the editors September 12, 1988; accepted for publication (in revised form) March 13, 1991.

† Department of Management and System Science, The Science University of Tokyo, Suwa College, 5000-1 Toyohira, Chino 391-02, Nagano, Japan.

parameters is to apply the *gradient method* in a formal way; setting the initial estimate \hat{a}_0 of the unknown parameter, the iterative scheme is given by

$$(1.0) \quad \hat{a}_{i+1} = \hat{a}_i + \lambda \times (\text{the first variation of the likelihood functional} \\ \text{w.r.t. the parameter } a)|_{a=\hat{a}_i} \quad \text{for } i = 0, 1, 2, \dots,$$

where $0 < \lambda < 1$.

However, the first variation of the likelihood functional (the second term of the right-hand side of (4.34) in § 4) does not have a regularity property with respect to the spatial variables; therefore, the solution \hat{a}_{i+1} of (1.0) does not belong to $C^2(G)$, even if the initial value \hat{a}_0 is in $C^2(G)$. (The regularity property is required for proving the consistency property of the MLE, as shown in [3].) Here, we use the regularization technique, which was proposed by Kravaris and Seinfeld [11] for deterministic systems. In [11] they showed that the regularization method plays an important role for achieving numerical stability. In this paper, the regularization method is used to guarantee the regularity property of the first variation of the (regularized) likelihood functional with respect to the parameter, as will be shown in Theorem 4.2 in § 4. Based on the first variation of the regularized likelihood functional, we can construct a feasible algorithm for generating the MLE that is similar to (1.0).

The following is a brief summary of this paper. Section 2 contains the relevant background results concerning the mathematical models of the considered systems and the consistency property of the MLE studied in [3]. In § 3, introducing the regularization technique to the likelihood functional, the consistency property of the regularized MLE (RMLE) is proved. In § 4 the necessary conditions for the RMLE are derived using the “adjoint equation” [12]. In § 5 an iterative algorithm to find a solution that satisfies the necessary conditions for the RMLE is proposed and the convergence of the generated sequence to the RMLE is shown. In the final section, we demonstrate some results of digital simulation experiments.

Let $G \subset R^n$ be a bounded open domain with smooth boundary $\Gamma = \partial G$. We define

$$(1.1) \quad V = H_0^1(G) \subset H = L^2(G) \subset V' = H^{-1}(G)$$

and denote the norm and inner product in H by $|\cdot|$ and (\cdot, \cdot) , respectively. We set¹

$$(1.2a) \quad A(a^0) \in \mathcal{L}(V; V')$$

and, for all $\phi_1, \phi_2 \in H$,

$$(1.2b) \quad \langle A(a^0)\phi_1, \phi_2 \rangle = \sum_{i=1}^n \int_G a^0(x) \frac{\partial \phi_1}{\partial x_i} \frac{\partial \phi_2}{\partial x_i} dx,$$

where a^0 is a function with values in R^1 and

$$(C1) \quad 0 < \alpha < \alpha^0 < \beta, \quad \forall x \in G,$$

$$(C2) \quad a^0 \in C^2(G).$$

From (C2), we find that operator $A(a^0)$ with range restricted to H defines an unbounded operator that is still denoted by $A(a^0)$, with the domain $\mathcal{D}(A) = H_0^1(G) \cap H^2(G)$. $(\Omega, \mathcal{F}, \mathcal{P})$ is a complete probability space and (\mathcal{F}_t) is an increasing family of sub- σ -algebras of \mathcal{F} . Let $w(t)$ be an H -valued Brownian motion process with the incremental covariance $Q \in \mathcal{L}_1(H; H)$, where $\mathcal{L}_1(\cdot, \cdot)$ denotes the class of

¹ Given X and Y Banach spaces, we denote by $\mathcal{L}(X, Y)$ the space of bounded linear operators from X into Y .

trace operators and v is an m -dimensional standard Brownian motion process independent of w .

Remark 1.1. It is possible to treat a more general elliptic partial differential operator than $A(a^0)$ defined by (1.2). However, to focus on our idea, we consider only this simple equation (1.2).

2. System model and MLE. In this section, we present the results of the MLE for an infinite-dimensional parameter in stochastic diffusion equations derived in [3]. Consider the following stochastic diffusion equation with the m -dimensional observation mechanism:

$$(2.1a) \quad (u(t), \phi) + \int_0^t \langle A(a^0)u(s), \phi \rangle ds = (u_0, \phi) + (w(t), \phi), \quad \forall \phi \in V,$$

$$(2.1b) \quad y(t) = \int_0^t Bu(s) ds + v(t),$$

where $B \in \mathcal{L}(H; R^m)$.

THEOREM 2.1 (see [6, Thm. 5.1, p. 189], [10], [16, Thm. 2.1, p. 48]). *Under (C1), (C2), and*

$$(C3) \quad u_0 \in L^2(\Omega; H) \text{ (zero-mean Gaussian)},^2$$

there exists a unique solution of (2.1a) such that

$$(2.2) \quad u \in L^2(\Omega; C(0, T; H) \cap L^2(0, T; V)).$$

LEMMA 2.1 (see [3, Lemma 2.1], [15, Cor. 2.13, p. 924], [18, Thm. 1, p. 252]). *We assume that the system has reached the stationary state. In the steady state, the filter equation is given by*

$$(2.3) \quad (\hat{u}(t, a^0), \phi) + \int_0^t \langle A(a^0)\hat{u}(s, a^0), \phi \rangle ds = \int_0^t \langle P(a^0)B^* dz(s; a^0), \phi \rangle, \quad \forall \phi \in V,$$

where $\hat{u}(t, a^0) = E\{u(t, a^0) | \mathcal{Y}_t\}$, $\mathcal{Y}_t = \sigma\{y(s); 0 \leq s \leq t\}$ and $z(t; a^0)$ is defined by

$$(2.4) \quad z(t; a^0) = y(t) - \int_0^t B\hat{u}(s, a^0) ds,$$

$P(a^0)$ is a unique solution of

$$(2.5) \quad \begin{aligned} & -\langle A(a^0)P(a^0)\phi_1, \phi_2 \rangle - \langle A^*(a^0)\phi_1, P(a^0)\phi_2 \rangle + (Q\phi_1, \phi_2) \\ & - (P(a^0)B^*BP(a^0)\phi_1, \phi_2) = 0, \quad \forall \phi_1, \phi_2 \in V. \end{aligned}$$

Remark 2.1. The stationarity assumption in Lemma 2.1 was verified in [3] by assuming the initial condition u_0 as $u_0 = \int_{-\infty}^0 T_\tau dw(\tau)$, where T_t is an exponentially stable semigroup generated by $A(a^0)$ [14, Thm. 5.5, Chap. 5]. From [12, Thm. 6.1, p. 171], for (2.5) we can show that

$$(2.6) \quad P(a^0) \in \mathcal{L}_2(H; V') \cap \mathcal{L}_2(V; H).$$

LEMMA 2.2 (see [3, Lemma 2.3]). *For any a and b that satisfy (C1) and (C2), there exists a constant C independent of a and b such that*

$$(2.7) \quad |P(a)|_{\mathcal{L}_2(V'; H)} + |P(a)|_{\mathcal{L}_2(H; V)} \leq C,$$

$$(2.8) \quad |(P(a) - P(b))B^*|^2_{\mathcal{L}(R^m; H)} \leq C|a - b|^2,$$

² Without Gaussian assumption, this theorem holds.

where $\mathcal{L}_2(\cdot, \cdot)$ denotes the Hilbert-Schmidt class of operators, and $|\cdot|_{\mathcal{L}_2}^2$ denotes the Hilbert-Schmidt norm.

LEMMA 2.3 (see [3, Thm. 3.2]). Let Θ be

$$(2.9) \quad \Theta = \{\theta \mid \theta \text{ satisfies conditions (C1) and (C2)}\}.$$

Assuming that

$$(C4) \quad B \in \mathcal{L}(V'; R^m),$$

we have

$$(2.10) \quad \lim_{T \rightarrow \infty} \sup_{a \in \Theta} \frac{1}{T} \int_0^T (B\hat{u}(t, a))^* dv(t) = 0 \quad \text{a.s.}$$

Using Lemma 2.3, the following consistency property can be proved in [3, Thm. 3.3].

THEOREM 2.2. Let M be the measure null-set outside that Lemma 2.3 holds. For each $\omega \notin M$, letting \hat{a}_T be the maximum likelihood estimate of a^0 (true value), we have the following strong consistency property:

$$(2.11) \quad \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T |B(\hat{u}(t, \hat{a}_T) - u(t, a^0))|_{R^m}^2 dt = \text{tr} [B\{P(a^0)B^*\}].$$

In Theorem 3.2 of the next section, we prove the consistency property of the RMLE, which is analogous to the proof of Theorem 2.2.

3. Maximum likelihood estimate by regularization. In this section, we reformulate the MLE by using the regularization technique proposed by Kravaris [11] in deterministic systems. The main purpose for applying this regularization technique to the likelihood estimate is to construct a feasible algorithm for generating the MLE, which will be discussed in § 5. Here, we show the consistency property for the regularized MLE. The cylindrical measure μ_y induced by y is absolutely continuous with respect to μ_v induced by v . From Liptser and Shiryaev [13, Thm. 7.13, p. 261], the Radon-Nikodym derivative is given by

$$(3.1) \quad \frac{d\mu_y(a)}{d\mu_v} = \exp \left\{ \int_0^T (B\hat{u}(t, x), dy(t))_{R^m} - \frac{1}{2} \int_0^T |B\hat{u}(t, a)|_{R^m}^2 dt \right\},$$

and the original likelihood functional is given by

$$(3.2) \quad L(T, y, a) = \ln \frac{d\mu_y(a)}{d\mu_v}.$$

To regularize the estimate of a^0 , instead of Θ , we introduce a regular space such that

$$(3.3) \quad \Theta_R = \left\{ a \in H^s(G) \mid 0 < \alpha \leq a(x) \leq \beta, \quad \forall x \in G, \quad \text{for } s > \frac{n}{2} + 2 \right\},$$

where s is an integer. From $s > n/2 + 2$ we find that the space Θ_R is densely imbedded in C^2 (see, e.g., [1, Thm. 5.4, p. 97]), i.e.,

$$(3.4) \quad \Theta_R \subset C^2.$$

Now we add the following regularization functional to (3.2): $-\gamma(T)\|a\|_{H^s}^2$; therefore, the regularized cost becomes

$$(3.5) \quad L_\gamma(T, y, a) = L(T, y, a) - \gamma(T)\|a\|_{H^s}^2,$$

where

$$(3.6) \quad \gamma(T) > 0 \text{ for } 0 < T < \infty$$

and

$$(3.7) \quad \lim_{T \rightarrow \infty} \frac{\gamma(T)}{T} = 0.$$

In the rest of this paper, all conditions and assumptions stated before are assumed to hold.

THEOREM 3.1. *There exists at least one RMLE $\hat{a}_\gamma^T \in \Theta_R$ such that*

$$L_\gamma(T, y, \hat{a}_\gamma^T) \geq L_\gamma(T, y, a), \quad \forall a \in \Theta_R.$$

Proof. See Theorem 3.1 of [3, p. 1069] for the proof.

THEOREM 3.2. *For $\hat{a}_\gamma^T = \arg \max_{a \in \Theta_R} L_\gamma$, we have the following strong consistency property:*

$$(3.8) \quad \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T |B(\hat{u}(t, \hat{a}_\gamma^T) - u(t, a^0))|_{R^m}^2 dt = \text{tr}[BP(a^0)B^*], \quad \text{a.s.}$$

Proof. We define

$$(3.9) \quad \tilde{\mu}_y(a) = \exp\{-\gamma(T)\|a\|_{H^s}^2\} \mu_y(a).$$

It follows from [13, Thm. 7.13, p. 261] that

$$\begin{aligned} \frac{d\tilde{\mu}_y(a)}{d\tilde{\mu}_y(a^0)} &= \frac{d\tilde{\mu}_y(a)}{d\mu_v} \frac{d\mu_v}{d\tilde{\mu}_y(a^0)} \\ &= \exp\left\{-\frac{1}{2} \int_0^T |B(u(t, a^0) - \tilde{u}(t, a))|_{R^m}^2 dt \right. \\ &\quad \left. + \frac{1}{2} \int_0^T |B(u(t, a^0) - \hat{u}(t, a^0))|_{R^m}^2 dt \right\} \\ &\quad \cdot \exp\left\{\int_0^T (B(\hat{u}(t, a) - \hat{u}(t, a^0)), dv(t))_{R^m}\right\} \\ &\quad \cdot \exp\{-\gamma(T)(\|a\|_{H^s}^2 - \|a^0\|_{H^s}^2)\}. \end{aligned} \quad (3.10)$$

Now, we find that the regularized MLE \hat{a}_γ^T satisfies the following relation:

$$(3.11) \quad \frac{d\tilde{\mu}_y(\hat{a}_\gamma^T)}{d\tilde{\mu}_y(a^0)} \geq 1.$$

Noting that all results of Lemma 2.3 are available for Θ_R , we can derive

$$(3.12) \quad \lim_{T \rightarrow \infty} \sup_{a \in \Theta_R} \frac{1}{T} \int_0^T (B(\hat{u}(t, a) - \hat{u}(t, a^0)), dv(t))_{R^m} = 0 \quad \text{a.s.}$$

Defining

$$(3.13) \quad I(T, a) = \int_0^T |B(u(t, a^0) - \hat{u}(t, a))|_{R^m}^2 dt,$$

from [3, Thm. 3.3] and [17, Thm. 1, p. 655], we have

$$(3.14) \quad \lim_{T \rightarrow \infty} \inf_{a \in \Theta_R} \left\{ \frac{1}{T} I(T, a) \right\} \geq \text{tr}[BP(a^0)B^*].$$

Consequently, from (3.14), we have

$$(3.15) \quad \lim_{T \rightarrow \infty} \frac{1}{T} \{I(T, \hat{a}_\gamma^T) - I(T, a^0)\} \geq 0 \quad \text{a.s.}$$

It follows from (3.7) that

$$(3.16) \quad \lim_{T \rightarrow \infty} \frac{1}{T} \gamma(T) \left\{ \|a^0\|_{H^s}^2 - \inf_{a \in \Theta_R} \|a\|_{H^s}^2 \right\} = 0.$$

Hence, using (3.11), we can derive the following relation:

$$(3.17) \quad \begin{aligned} & \lim_{T \rightarrow \infty} \frac{1}{T} \left\{ \int_0^T (B(\hat{u}(t, \hat{a}_\gamma^T) - \hat{u}(t, a^0)), dv(t))_{R^m} - \gamma(T) \{ \|a^0\|_{H^s}^2 - \|\hat{a}_\gamma^T\|_{H^s}^2 \} \right\} \\ & \geq \lim_{T \rightarrow \infty} \frac{1}{2T} \{I(T, \hat{a}_\gamma^T) - I(T, a^0)\} \geq 0. \end{aligned}$$

Thus, (3.8) can be derived.

Remark 3.1. From Theorem 3.2, we also have

$$\lim_{T \rightarrow \infty} \text{tr} [BP(\hat{a}_\gamma^T)B^*] = \text{tr} [BP(a^0)B^*] \quad \text{a.s.}$$

4. Necessary conditions for the RMLE. Before deriving the necessary conditions for the RMLE, we transform the measure \mathcal{P} to $\tilde{\mathcal{P}}$ such that

$$(4.1) \quad d\mathcal{P} = q d\tilde{\mathcal{P}},$$

where

$$(4.2) \quad q = \exp \left\{ \int_0^T (Bu(t), dy(t))_{R^m} - \frac{1}{2} \int_0^T |Bu(t)|_{R^m}^2 \right\}.$$

Hence, we find that y is a Wiener process with respect to the measure $\tilde{\mathcal{P}}$.

First, we discuss the first variation of $L_\gamma(T, y, a)$ with respect to the parameter a . Defining

$$(4.3) \quad (e(t), \phi) = \lim_{\varepsilon \rightarrow 0} (\hat{u}(t, a + \varepsilon \delta a) - \hat{u}(t, a), \phi) / \varepsilon,$$

for $a, a + \varepsilon \delta a \in \Theta_R$, we obtain

$$(4.4) \quad \begin{aligned} & (e(t), \phi) + \int_0^t \langle (A(a) + \mathbf{P}(a)B^*B) e(s), \phi \rangle ds \\ & = \int_0^t (\tilde{\mathbf{P}}(\delta a)B^* dy(s), \phi) - \int_0^t \langle (\tilde{A}(\delta a) + \tilde{\mathbf{P}}(\delta a)B^*B) \hat{u}(s, a), \phi \rangle ds, \quad \forall \phi \in V, \end{aligned}$$

where, for $\phi_1, \phi_2 \in V$,

$$(4.5) \quad \langle \tilde{A}(\delta a) \phi_1, \phi_2 \rangle = \lim_{\varepsilon \rightarrow 0} \langle (A(a + \varepsilon \delta a) - A(a)) \phi_1, \phi_2 \rangle / \varepsilon = \sum_{i=1}^n \left(\delta a \frac{\partial \phi_1}{\partial x_i} \frac{\partial \phi_2}{\partial x_i} \right)$$

and

$$(4.6) \quad (\tilde{\mathbf{P}}(\delta a) \phi_1, \phi_2) = \lim_{\varepsilon \rightarrow 0} ((\mathbf{P}(a + \varepsilon \delta a) - \mathbf{P}(a)) \phi_1, \phi_2) / \varepsilon.$$

Using (4.6), (2.5) becomes

$$(4.7) \quad \begin{aligned} & -\langle A(a) \tilde{\mathbf{P}}(\delta a) \phi_1, \phi_2 \rangle - \langle A^*(a) \phi_1, \tilde{\mathbf{P}}(\delta a) \phi_2 \rangle \\ & - ((\tilde{\mathbf{P}}(\delta a)B^*B\mathbf{P}(a) + \mathbf{P}(a)B^*B\tilde{\mathbf{P}}(\delta a)) \phi_1, \phi_2) \\ & = \langle \tilde{A}(\delta a) \mathbf{P}(a) \phi_1, \phi_2 \rangle + \langle \tilde{A}^*(\delta a) \phi_1, \mathbf{P}(a) \phi_2 \rangle. \end{aligned}$$

Furthermore, from (2.6), we also find that (4.7) has the following unique solution:

$$(4.8) \quad \tilde{\mathbf{P}}(\delta a) \in \mathcal{L}(V'; H) \cap \mathcal{L}(H; V).$$

It can be easily shown that

$$(4.9) \quad e \in L^2(\Omega, \tilde{\mathcal{F}}; C(0, T; H) \cap L^2(0, T; V)).$$

Then the first variation of the cost $L_\gamma(T, y, a)$ becomes

$$(4.10) \quad \begin{aligned} \delta L_\gamma(T, y, a) &= \lim_{\varepsilon \rightarrow 0} \frac{L_\gamma(T, y, a + \varepsilon \delta a) - L_\gamma(T, y, a)}{\varepsilon} \\ &= \int_0^T (Be(t), dy(t))_{R^m} - \int_0^T (Be(t), B\hat{u}(t, a))_{R^m} dt - 2\gamma(T)(\delta a, a)_{H^s}. \end{aligned}$$

To derive the necessary conditions for the RMLE, we must introduce the adjoint equation. To do this, (4.4) is transformed into the pathwise form. We define

$$(4.11) \quad \tilde{e}(t) = e(t) - \tilde{\mathbf{P}}(\delta a)B^*y(t).$$

Using (4.11), (4.4) can be rewritten as

$$(4.12) \quad \begin{aligned} &\left\langle \frac{d\tilde{e}(t)}{dt} + (A(a) + \mathbf{P}(a)B^*B)(\tilde{e}(t) + \tilde{\mathbf{P}}(\delta a)B^*y(t)), \phi \right\rangle \\ &= -\langle (\tilde{A}(\delta a) + \tilde{\mathbf{P}}(\delta a)B^*B)\hat{u}(t, a), \phi \rangle, \quad \forall \phi \in V, \\ &\tilde{e}(0) = 0. \end{aligned}$$

PROPOSITION 4.1. Equation (4.12) has a unique solution such that

$$(4.13) \quad \frac{d\tilde{e}}{dt} \in L^2(0, T; V') \quad \text{and} \quad \tilde{e} \in L^2(0, T; V), \quad \text{a.s. } \tilde{\mathcal{F}}.$$

Proof. Noting that $\hat{u}(\cdot, a) \in C(0, T; H) \cap L^2(0, T; V)$ almost surely $\tilde{\mathcal{F}}$ and $y \in C(0, T; R^m)$, almost surely $\tilde{\mathcal{F}}$, we find that

$$(4.14) \quad (A(a) + \mathbf{P}(a)B^*B)\tilde{\mathbf{P}}(\delta a)B^*y \in L^2(0, T; V'), \quad \text{a.s. } \tilde{\mathcal{F}}$$

and

$$(4.15) \quad (\tilde{A}(\delta a) + \tilde{\mathbf{P}}(\delta a)B^*B)\hat{u} \in L^2(0, T; V'), \quad \text{a.s. } \tilde{\mathcal{F}}.$$

From [12, p. 102], (4.13) can be derived.

PROPOSITION 4.2. Introducing the following adjoint equation for $\tilde{e}(t)$ -equation:

$$(4.16) \quad \begin{aligned} &\left\langle \frac{d\tilde{p}(t, a)}{dt} - (A^*(a) + B^*B\mathbf{P}(a))(\tilde{p}(t, a) + B^*(y(T) - y(t))), \phi \right\rangle \\ &= (B^*B\hat{u}(t, a), \phi), \quad \forall \phi \in V, \\ &\tilde{p}(T, a) = 0, \end{aligned}$$

(4.16) has a unique solution such that

$$(4.17) \quad \frac{d\tilde{p}}{dt} \in L^2(0, T; V') \quad \text{and} \quad \tilde{p} \in L^2(0, T; V), \quad \text{a.s. } \tilde{\mathcal{F}}.$$

Proof. Noting that $B^* \in \mathcal{L}(R^m; V)$, we have

$$(4.18) \quad (A^*(a) + B^*B\mathbf{P}(a))B^*(y(T) - y(\cdot)) + B^*B\hat{u}(\cdot, a) \in L^2(0, T; V') \quad \text{a.s. } \tilde{\mathcal{F}}.$$

From [12, p. 102], we find that (4.16) has a unique solution, as shown in (4.17).

PROPOSITION 4.3. *Introducing the adjoint equation for $\mathbf{P}(a)$ -equation, namely*

$$\begin{aligned}
 & -\langle A^*(a)\mathbf{R}(a)\phi_1, \phi_2 \rangle - \langle A(a)\phi_1, \mathbf{R}(a)\phi_2 \rangle \\
 & -((\mathbf{P}(a)B^*B\mathbf{R}(a) + \mathbf{R}(a)B^*B\mathbf{P}(a))\phi_1, \phi_2) \\
 & = \frac{1}{2} \left(\left(\int_0^T B^*y(t) \otimes B^* dy(t) - \int_0^T B^*y(t) \otimes \frac{d\tilde{p}(t, a)}{dt} dt \right. \right. \\
 (4.19) \quad & \left. \left. - \int_0^T B^*B\hat{u}(t, a) \otimes (\tilde{p}(t, a) + B^*(y(T) - y(t))) dt \right) \phi_1, \phi_2 \right) \\
 & + \frac{1}{2} \left(\phi_1, \left(\int_0^T B^*y(t) \otimes B^* dy(t) - \int_0^T B^*y(t) \otimes \frac{d\tilde{p}(t, a)}{dt} dt \right. \right. \\
 & \left. \left. - \int_0^T B^*B\hat{u}(t, a) \otimes (\tilde{p}(t, a) + B^*(y(T) - y(t))) dt \right) \phi_2 \right), \quad \forall \phi_1, \phi_2 \in V,
 \end{aligned}$$

where $\phi_1 \otimes \phi_2 = \phi_1(\phi_2, \cdot)$, (4.19) has a unique solution in

$$(4.20) \quad \mathbf{R} \in \mathcal{L}_2(V'; H) \cap \mathcal{L}_2(H; V).$$

Proof. To show the existence of the solution to (4.19), it is sufficient to show that

$$\begin{aligned}
 (4.21) \quad & \left[\int_0^T B^*y(t) \otimes B^* dy(t) - \int_0^T B^*B\hat{u}(t, a) \otimes (\tilde{p}(t, a) + B^*(y(T) - y(t))) dt, \mathbf{R}(a) \right] \\
 & \leq \text{const } |\mathbf{R}(a)|_{\mathcal{L}_2(H; H)} \quad \text{a.s. } \tilde{\mathcal{P}}
 \end{aligned}$$

and

$$(4.22) \quad \left[\int_0^T B^*y(t) \otimes \frac{d\tilde{p}(t, a)}{dt} dt, \mathbf{R}(a) \right] \leq \text{const } |\mathbf{R}(a)|_{\mathcal{L}_2(V'; H)}, \quad \text{a.s. } \tilde{\mathcal{P}},$$

where $[\cdot, \cdot]$ denotes the Hilbert-Schmidt inner product. These estimates follow from the following properties:

$$(4.23) \quad \int_0^T B^*y(t) \otimes B^* dy(t) \in \mathcal{L}_2(V'; V), \quad \text{a.s. } \tilde{\mathcal{P}},$$

$$(4.24) \quad \int_0^T B^*B\hat{u}(t, a) \otimes (\tilde{p}(t, a) + B^*(y(T) - y(t))) dt \in \mathcal{L}_2(V'; V), \quad \text{a.s. } \tilde{\mathcal{P}},$$

and

$$(4.25) \quad \int_0^T B^*y(t) \otimes \frac{d\tilde{p}(t, a)}{dt} dt \in \mathcal{L}_2(V; V), \quad \text{a.s. } \tilde{\mathcal{P}}.$$

Furthermore, it follows from $\mathbf{P} \in \mathcal{L}_2(V'; H) \cap \mathcal{L}_2(H; V)$ and [18] that (4.19) has a unique solution in (4.20).

Now we will derive the necessary conditions for the RMLE.

THEOREM 4.1. *The first variation of $L_\gamma(T, y, a)$ is given by*

$$\begin{aligned}
 (4.26) \quad \delta L_\gamma(T, y, a) = & \int_0^T \left(- \sum_{i=1}^n \frac{\partial \hat{u}(t, a)}{\partial x_i} \frac{\partial (\tilde{p}(t, a) + B^*(y(T) - y(t)))}{\partial x_i}, \delta a \right) dt \\
 & + \left(2 \sum_{i=1}^n \int_{G_z} \left(\frac{\partial R(a, x, z)}{\partial x_i} \frac{\partial P(a, z, x)}{\partial x_i} \right) dz, \delta a \right) - 2\gamma(T)(a, \delta a)_{H^s},
 \end{aligned}$$

where $R(a, \cdot, \cdot)$ and $P(a, \cdot, \cdot)$ are the kernels of $\mathbf{R}(a)$ and $\mathbf{P}(a)$, respectively.

Proof. From (4.12) and (4.16), we have

$$\begin{aligned}
 (\tilde{e}(T), \tilde{p}(T, a)) - (\tilde{e}(0), \tilde{p}(0, a)) &= \int_0^T \left\{ \left\langle \frac{d\tilde{e}(t)}{dt}, \tilde{p}(t, a) \right\rangle + \left\langle \tilde{e}(t), \frac{d\tilde{p}(t, a)}{dt} \right\rangle \right\} dt \\
 &= \int_0^T \{ -\langle (A(a) + \mathbf{P}(a)B^*B)(\tilde{e}(t) + \tilde{\mathbf{P}}(\delta a)B^*y(t)), \tilde{p}(t, a) \rangle \\
 &\quad - \langle (\tilde{A}(\delta a) + \tilde{\mathbf{P}}(\delta a)B^*B)\hat{u}(t, a), \tilde{p}(t, a) \rangle \\
 &\quad + \langle \tilde{e}(t), (A^*(a) + B^*B\mathbf{P}(a))(\tilde{p}(t, a) + B^*(y(T) - y(t))) \rangle \\
 &\quad + \langle \tilde{e}(t), B^*B\hat{u}(t, a) \rangle \} dt.
 \end{aligned}$$

Noting that $\tilde{e}(0) = \tilde{p}(T, a) = 0$, the above equation becomes

$$\begin{aligned}
 & - \int_0^T \{ \langle (A(a) + \mathbf{P}(a)B^*B)\tilde{\mathbf{P}}(\delta a)B^*y(t) + (\tilde{A}(\delta a) + \tilde{\mathbf{P}}(\delta a)B^*B)\hat{u}(t, a), \tilde{p}(t, a) \rangle \} dt \\
 (4.27) \quad & + \int_0^T \langle \tilde{e}(t), (A^*(a) + B^*B\mathbf{P}(a))B^*(y(T) - y(t)) \rangle dt \\
 & + \int_0^T \langle \tilde{e}(t), B^*B\hat{u}(t, a) \rangle dt = 0.
 \end{aligned}$$

From (4.12), the second term of the left-hand side of (4.27) becomes

$$\begin{aligned}
 & \int_0^T \langle (A(a) + \mathbf{P}(a)B^*B)\tilde{e}(t), B^*(y(T) - y(t)) \rangle dt \\
 &= \int_0^T \left\langle -\frac{d\tilde{e}(t)}{dt}, B^*(y(T) - y(t)) \right\rangle dt \\
 &\quad - \int_0^T \langle (A(a) + \mathbf{P}(a)B^*B)\tilde{\mathbf{P}}(\delta a)B^*y(t) \\
 &\quad + (\tilde{A}(\delta a) + \tilde{\mathbf{P}}(\delta a)B^*B)\hat{u}(t, a), B^*(y(T) - y(t)) \rangle dt \\
 &= I_1 + I_2,
 \end{aligned}$$

for example. It is easy to show that

$$\begin{aligned}
 I_1 &= - \int_0^T \left\langle \frac{d\tilde{e}(t)}{dt}, B^*y(T) \right\rangle dt + \int_0^T \left\langle \frac{d\tilde{e}(t)}{dt}, B^*y(t) \right\rangle dt \\
 (4.28) \quad &= -(\tilde{e}(T), B^*y(T)) + (\tilde{e}(0), B^*y(T)) + \int_0^T \left\langle \frac{d\tilde{e}(t)}{dt}, B^*y(t) \right\rangle dt.
 \end{aligned}$$

Noting from (4.11) that $\tilde{e}(t)$ is \mathcal{Y}_t -measurable, the third term of the right-hand side of (4.28) can be represented as the stochastic integral with respect to y . From $\tilde{e}(0) = 0$, we get

$$I_1 = - \int_0^T (\tilde{e}(t), B^* dy(t)) = - \int_0^T (Be(t), dy(t))_{\mathbb{R}^m} + \int_0^T (\tilde{\mathbf{P}}(\delta a)B^*y(t), B^* dy(t)).$$

Hence, from (4.27), we have

$$\begin{aligned}
 (4.29) \quad & \int_0^T \langle (A(a) + \mathbf{P}(a)B^*B)\tilde{\mathbf{P}}(\delta a)B^*y(t), \tilde{p}(t, a) + B^*(y(T) - y(t)) \rangle dt \\
 & + \int_0^T \langle (\tilde{A}(\delta a) + \tilde{\mathbf{P}}(\delta a)B^*B)\tilde{u}(t, a), \tilde{p}(t, a) + B^*(y(T) - y(t)) \rangle dt \\
 & + \int_0^T (Be(t), dy(t))_{\mathbb{R}^m} - \int_0^T (\tilde{\mathbf{P}}(\delta a)B^*y(t), B^*dy(t)) \\
 & - \int_0^T \langle e(t) - \tilde{\mathbf{P}}(\delta a)B^*y(t), B^*B\tilde{u}(t, a) \rangle dt = 0.
 \end{aligned}$$

From (4.16), the first term of the left-hand side of (4.29) becomes

$$\begin{aligned}
 & \int_0^T \langle \tilde{\mathbf{P}}(\delta a)B^*y(t), (A^*(a) + B^*B\mathbf{P}(a))(\tilde{p}(t, a) + B^*(y(T) - y(t))) \rangle dt \\
 & = \int_0^T \left\langle \tilde{\mathbf{P}}(\delta a)B^*y(t), \frac{d\tilde{p}(t, a)}{dt} - B^*B\tilde{u}(t, a) \right\rangle dt.
 \end{aligned}$$

Combining all the estimates, we have

$$\begin{aligned}
 (4.30) \quad & \int_0^T (Be(t), dy(t))_{\mathbb{R}^m} - \int_0^T (Be(t), B\tilde{u}(t, a))_{\mathbb{R}^m} dt \\
 & = - \int_0^T \langle (\tilde{A}(\delta a) + \tilde{\mathbf{P}}(\delta a)B^*B)\tilde{u}(t, a), \tilde{p}(t, a) + B^*(y(T) - y(t)) \rangle dt \\
 & \quad - \int_0^T \left\langle \tilde{\mathbf{P}}(\delta a)B^*y(t), \frac{d\tilde{p}(t, a)}{dt} \right\rangle dt + \int_0^T (\tilde{\mathbf{P}}(\delta a)B^*y(t), B^*dy(t)).
 \end{aligned}$$

Furthermore, from (4.7) and (4.19), we have

$$\begin{aligned}
 (4.31a) \quad & -[A^*(a)\mathbf{R}(a), \tilde{\mathbf{P}}(\delta a)] - [A(a)\tilde{\mathbf{P}}(\delta a), \mathbf{R}(a)] - [\mathbf{P}(a)B^*B\mathbf{R}(a) + \mathbf{R}(a)B^*B\mathbf{P}(a), \tilde{\mathbf{P}}(\delta a)] \\
 & = \left[\int_0^T B^*y(t) \otimes B^*dy(t) - \int_0^T \left\{ B^*y(t) \otimes \frac{d\tilde{p}(t, a)}{dt} \right. \right. \\
 & \quad \left. \left. - B^*B\tilde{u}(t, a) \otimes (\tilde{p}(t, a) + B^*(y(T) - y(t))) \right\} dt, \tilde{\mathbf{P}}(\delta a) \right] \\
 & = \int_0^T (\tilde{\mathbf{P}}(\delta a)B^*y(t), B^*dy(t)) - \int_0^T \left\langle \tilde{\mathbf{P}}(\delta a)B^*y(t), \frac{d\tilde{p}(t, a)}{dt} \right\rangle dt \\
 & \quad - \int_0^T (\tilde{\mathbf{P}}(\delta a)B^*B\tilde{u}(t, a), \tilde{p}(t, a) + B^*(y(T) - y(t))) dt
 \end{aligned}$$

and

$$\begin{aligned}
 (4.31b) \quad & -[A^*(a)\mathbf{R}(a), \tilde{\mathbf{P}}(\delta a)] - [A(a)\tilde{\mathbf{P}}(\delta a), \mathbf{R}(a)] - [\mathbf{P}(a)B^*B\mathbf{R}(a) + \mathbf{R}(a)B^*B\mathbf{P}(a), \tilde{\mathbf{P}}(\delta a)] \\
 & = 2[\tilde{A}(\delta a)\mathbf{P}(a), \mathbf{R}(a)] = 2\left(\left(\int_G \sum_{i=1}^n \frac{\partial R(a, x, z)}{\partial x_i} \frac{\partial P(a, z, x)}{\partial x_i} dz\right), \delta a\right).
 \end{aligned}$$

Then, it follows from (4.31a) and (4.31b) that

$$\begin{aligned}
 & \int_0^T (\tilde{\mathbf{P}}(\delta a) B^* y(t), B^* dy(t)) - \int_0^T \left\langle \tilde{\mathbf{P}}(\delta a) B^* y(t), \frac{d\tilde{p}(t, a)}{dt} \right\rangle dt \\
 & - \int_0^T (\tilde{\mathbf{P}}(\delta a) B^* \hat{B} \hat{u}(t, a), \tilde{p}(t, a) + B^*(y(T) - y(t))) dt \\
 & = 2 \left(\left(\int_G \sum_{i=1}^n \frac{\partial R(a, x, z)}{\partial x_i} \frac{\partial P(a, z, x)}{\partial x_i} dz \right), \delta a \right).
 \end{aligned}
 \tag{4.32}$$

Hence, combining (4.10), (4.30), and (4.32), (4.26) can be derived.

Now we introduce the following operator $\Lambda \in \mathcal{L}(H^s; (H^s)')$ such that

$$\langle \Lambda \phi_1, \phi_2 \rangle = \sum_{i=1}^n \left(\frac{\partial^s \phi_1}{\partial x_i^s}, \frac{\partial^s \phi_2}{\partial x_i^s} \right) + (\phi_1, \phi_2), \quad \text{for } \phi_1, \phi_2 \in H^s,$$

where $\langle \cdot, \cdot \rangle$ denotes the duality between H^s and $(H^s)'$.

Noting that the norm and the inner product in H^s are represented by

$$\|\phi\|_{H^s}^2 = (\phi, \phi)_{H^s} = \langle \Lambda \phi, \phi \rangle, \quad \text{for } \phi \in H^s,$$

from Theorem 4.1, the precise form of the necessary conditions can be derived.

THEOREM 4.2. *The necessary conditions of the RMLE are characterized by $\hat{u}(t, \hat{a}_\gamma^T)$, $\mathbf{P}(\hat{a}_\gamma^T)$, $\tilde{p}(t, \hat{a}_\gamma^T)$, $\mathbf{R}(\hat{a}_\gamma^T)$, and the variational inequality below:*

$$2\gamma(T) \langle \Lambda \hat{a}_\gamma^T, a - \hat{a}_\gamma^T \rangle \geq (f(\hat{a}_\gamma^T), a - \hat{a}_\gamma^T), \quad \forall a \in \Theta_R,
 \tag{4.33}$$

where $\hat{a}_\gamma^T \in \Theta_R$ is the RMLE and

$$\begin{aligned}
 f(\hat{a}_\gamma^T) = & \sum_{i=1}^n \left\{ 2 \int_G \frac{\partial R(\hat{a}_\gamma^T, x, z)}{\partial x_i} \frac{\partial P(\hat{a}_\gamma^T, z, x)}{\partial x_i} dz \right. \\
 & \left. - \int_0^T \frac{\partial \hat{u}(t, \hat{a}_\gamma^T)}{\partial x_i} \frac{\partial (\tilde{p}(t, \hat{a}_\gamma^T) + B^*(y(T) - y(t)))}{\partial x_i} dt \right\}.
 \end{aligned}
 \tag{4.34}$$

Moreover, there exists a solution \hat{a}_γ^T of the variational inequality (4.33). The uniqueness of the estimate \hat{a}_γ^T is not discussed here.

Remark 4.1. Assume that $|f(a_1) - f(a_2)|_{(H^s)'}^2 \leq |a_1 - a_2|^2$ for all a_1 and $a_2 \in \Theta_R$. Equation (4.33) has a unique solution. However, we cannot present the exact conditions for proving the uniqueness property here. In § 6 we will show three numerical examples that seem to guarantee the uniqueness property.

Proof. From the results of Theorem 4.1, (4.33) can easily be derived. The remaining task is to show the existence of the solution to the variational inequality (4.33). Applying Theorem 2.1 in [9, p. 24] for elliptic variational inequalities, the existence \hat{a}_γ^T in Θ_R can be proved, if all conditions stated in [9, Thm. 21] are satisfied. Noting that in (3.3) α and β are positive constants, the space Θ_R is evidently a closed convex set in H^s , and the operator Λ is coercive in H^s . Now we show that $f(\hat{a}_\gamma^T)$ is in $(H^s)'$, almost surely. From (2.7) and (4.20), we have

$$\begin{aligned}
 & \left\| \sum_{i=1}^n 2 \int_G \frac{\partial R(\hat{a}_\gamma^T, x, z)}{\partial x_i} \frac{\partial P(\hat{a}_\gamma^T, z, x)}{\partial x_i} dz \right\|_{(H^s)'} \\
 & \leq \text{const} \|\mathbf{R}(\hat{a}_\gamma^T)\|_{\mathcal{L}_2(H; H)}^2 \|\mathbf{P}(\hat{a}_\gamma^T)\|_{\mathcal{L}_2(H; H)}^2 < \infty \quad \text{a.s. } \tilde{\mathcal{P}}.
 \end{aligned}
 \tag{4.35}$$

Defining $\tilde{u}(t, \hat{a}_\gamma^T) = \hat{u}(t, \hat{a}_\gamma^T) - \mathbf{P}(\hat{a}_\gamma^T) B^* y(t)$ and applying the same technique used in the proof of Proposition 4.1, we also find that $\tilde{u}(\cdot, a) \in C(0, T; H)$ almost surely $\tilde{\mathcal{P}}$. Hence, from $y \in C(0, T; R^m)$, almost surely $\tilde{\mathcal{P}}$ and $\mathbf{P}(\hat{a}_\gamma^T) B^* \in \mathcal{L}(R^m; V)$, we have

$\hat{u}(\cdot, a) \in C(0, T; H)$ almost surely $\tilde{\mathcal{P}}$. (See the proof of Lemma 5.2 in § 5 for more details.) Furthermore, noting that from (4.17), $\tilde{p} \in C(0, T; H)$ almost surely $\tilde{\mathcal{P}}$, we have

$$(4.36) \quad \left\| \int_0^T \frac{\partial \hat{u}(t, \hat{a}_\gamma^T)}{\partial x_i} \frac{\partial (\tilde{p}(t, \hat{a}_\gamma^T) + B^*(y(T) - y(t)))}{\partial x_i} dt \right\|_{(h^*)}^2 \leq \text{const} \left(\sup_{0 \leq t \leq T} |\hat{u}(t, \hat{a}_\gamma^T)|^2 \right) \left(\sup_{0 \leq t \leq T} |\tilde{p}(t, \hat{a}_\gamma^T)|^2 \right) \left(\sup_{0 \leq t \leq T} |B^*(y(T) - y(t))|^2 \right) < \infty.$$

Hence, we find that $f(\hat{a}_\gamma^T) \in (H^s)'$, almost surely $\tilde{\mathcal{P}}$.

Remark 4.2. If the function $a \rightarrow L_\gamma(T, y, a)$ is strictly convex, the necessary conditions stated in Theorem 4.2 also become sufficient. However, it is generally impossible to prove the strict convexity property. The more important step is showing the uniqueness of the solution to the variational inequality (4.34) stated in Remark 4.1. To do this, we must calculate the second variation of $L_\gamma(T, y, a)$ with respect to a .

5. Iterative algorithm for the RMLE. In this section, we present an algorithm for generating the RMLE that satisfies the necessary conditions derived in § 4. The main difficulty in solving the elliptic variational inequality (4.33) is the nonlinearity of $f(a_\gamma^T)$. To circumvent this, the variational inequality (4.33) is approximated to the iterative form given by (5.4) below. From this approximation, we can construct the following iterative algorithm:

- (i) Take the initial value $\hat{a}_{i|i=0} = a \in \Theta_R \cap \{a \mid \|a\|_{H^s} \leq \text{const}\}$
- (ii) For \hat{a}_i , solve the filter equation³ and the Riccati equation:

$$(5.1a) \quad (\hat{u}^i(t), \phi) + \int_0^t \langle (A(\hat{a}_i) + \mathbf{P}(\hat{a}_i)B^*B)\hat{u}^i(s), \phi \rangle ds = (\mathbf{P}(\hat{a}_i)B^*y(t), \phi), \quad \forall \phi \text{ in } V,$$

$$(5.1b) \quad \begin{aligned} & -\langle A(\hat{a}_i)\mathbf{P}(\hat{a}_i)\phi_1, \phi_2 \rangle - \langle A^*(\hat{a}_i)\phi_1, \mathbf{P}(\hat{a}_i)\phi_2 \rangle + \langle Q\phi_1, \phi_2 \rangle \\ & - (\mathbf{P}(\hat{a}_i)B^*B\mathbf{P}(\hat{a}_i)\phi_1, \phi_2) = 0, \quad \forall \phi_1, \phi_2 \in V. \end{aligned}$$

- (iii) From \hat{a}_i , \hat{u}^i , and $\mathbf{P}(\hat{a}_i)$, solve the following adjoint equations:

$$(5.2a) \quad \begin{aligned} & \left\langle \frac{d\tilde{p}^i(t)}{dt} - (A^*(\hat{a}_i) + B^*B\mathbf{P}(\hat{a}_i))(\tilde{p}^i(t) + B^*(y(T) - y(t))), \phi \right\rangle \\ & = (B^*B\hat{u}^i(t), \phi), \quad \forall \phi \in V, \\ & \tilde{p}^i(T) = 0, \end{aligned}$$

and

$$(5.2b) \quad \begin{aligned} & -\langle A^*(\hat{a}_i)\mathbf{R}(\hat{a}_i)\phi_1, \phi_2 \rangle - \langle A(\hat{a}_i)\phi_1, \mathbf{R}(\hat{a}_i)\phi_2 \rangle \\ & - ((\mathbf{P}(\hat{a}_i)B^*B\mathbf{R}(\hat{a}_i) + \mathbf{R}(\hat{a}_i)B^*B\mathbf{P}(\hat{a}_i))\phi_1, \phi_2) \\ & = \frac{1}{2} \left(\left\{ \int_0^T B^*y(t) \otimes B^* dy(t) - \int_0^T B^*y(t) \otimes \frac{d\tilde{p}^i(t)}{dt} dt \right. \right. \\ & \quad \left. \left. - \int_0^T B^*B\hat{u}^i(t) \otimes (\tilde{p}^i(t) - B^*(y(T) - y(t))) dt \right\} \phi_1, \phi_2 \right) \\ & \quad + \frac{1}{2} \left(\phi_1, \left\{ \int_0^T B^*y(t) \otimes B^* dy(t) - \int_0^T B^*y(t) \otimes \frac{d\tilde{p}^i(t)}{dt} dt \right. \right. \\ & \quad \left. \left. - \int_0^T B^*B\hat{u}^i(t) \otimes (\tilde{p}^i(t) - B^*(y(T) - y(t))) dt \right\} \phi_2 \right), \quad \forall \phi_1, \phi_2 \in V. \end{aligned}$$

³ In the numerical experiments, it is convenient to transform the filter equation into the pathwise form as in (5.12) of Lemma 5.2, because u^i cannot be defined in the Itô sense for $i \geq 1$.

(iv) After calculating

$$(5.3) \quad f(\hat{a}_i) = \sum_{k=1}^n \left\{ 2 \int_G \frac{\partial R(\hat{a}_i, x, z)}{\partial x_k} \frac{\partial P(\hat{a}_i, z, x)}{\partial x_k} dz - \int_0^T \frac{\partial \hat{u}^i(t)}{\partial x_k} \frac{\partial (\tilde{p}^i(t) + B^*(y(T) - y(t)))}{\partial x_k} dt \right\},$$

the following elliptic variational inequality is solved:

$$(5.4) \quad 2\gamma(T) \langle \Lambda \hat{a}_{i+1}, a - \hat{a}_{i+1} \rangle \geq (f(\hat{a}_i), a - \hat{a}_{i+1}), \quad \forall a \in \Theta_R, \\ \hat{a}_{i+1} \in \Theta_R.$$

(v) Replacing \hat{a}_i by a newly obtained \hat{a}_{i+1} from (5.4), repeat steps (ii) through (v).

THEOREM 5.1. *Let \hat{a}_i be a sequence generated by the above algorithm. There exists a subsequence of \hat{a}_i (still denoted by \hat{a}_i) such that*

$$(5.5) \quad \hat{a}_i \rightarrow \hat{a}_\gamma^T \text{ weakly in } \Theta_R, \quad \text{a.s. } \tilde{\mathcal{P}},$$

where \hat{a}_γ^T is an RMLE.

Remark 5.1. The algorithm presented here is an off-line scheme. For each iteration step, we must solve two partial differential equations (5.1a)–(5.2a), the operator “Riccati equation” (5.1b), the “Lyapunov equation” (5.2b), and variational inequality (5.4). A typical program execution takes a long CPU time on the typical personal computer. At present, we cannot present a feasible on-line scheme.

The proof of Theorem 5.1 is not straightforward because the sequence \hat{a}_i depends on the whole value $y(t)$, $0 \leq t \leq T$ and then \hat{u}^i is no longer a Markov process; i.e., we cannot define the u^i -equation in the Itô sense. The mathematical justification of the \hat{u}^i -equation is given by using the pathwise form that was used for defining the \tilde{p} and \tilde{e} processes. (See Lemma 5.2 below.) Now we present the key lemmas for proving Theorem 5.1.

LEMMA 5.1. *Let \hat{a}_i be in a bounded set in Θ_R as shown in (i). Then we can extract a subsequence of \hat{a}_i , still denoted by \hat{a}_i such that $\hat{a}_i \rightarrow \tilde{a}$ weakly in H^s and strongly in H^{s-1} . Hence*

$$(5.6) \quad \mathbf{P}(\hat{a}_i)B^* \rightarrow \mathbf{P}(\tilde{a})B^* \text{ strongly in } \mathcal{L}(R^m; H), \quad \text{a.s. } \tilde{\mathcal{P}}.$$

Proof. The smoothness assumption for the domain G implies that the imbedding from H^s into H^{s-1} is compact [1, Chap. 6]. Because \hat{a}_i is in the bounded set in H^s , \hat{a}_i converges to \tilde{a} weakly in H^s and strongly in $H^{s-1} \subset L^2(G)$. For the algebraic Riccati equation (2.5) (see also (5.1b)), we have obtained the following estimates [3, Lemma 2.3]:

$$(5.7) \quad |\mathbf{P}(a)|_{\mathcal{L}_2(V; H)}^2 + |\mathbf{P}(a)|_{\mathcal{L}_2(H; V)}^2 \leq C_1$$

and

$$(5.8) \quad |(\mathbf{P}(a) - \mathbf{P}(b))B^*|_{\mathcal{L}(R^m; H)}^2 + |(\mathbf{P}(a) - \mathbf{P}(b))B^*|_{\mathcal{L}(R^m; H)}^2 \leq C_2|a - b|^2$$

for all $a, b \in \Theta_R$, where C_1 and C_2 are independent of a and b . Hence, by using the fact that \hat{a}_i converges to \tilde{a} strongly in H^{s-1} , (5.6) can be derived from (5.8).

LEMMA 5.2. *Let \hat{a}_i be the same sequence as in Lemma 5.1. Denoting*

$$(5.9) \quad \tilde{u}^i(t) = \hat{u}^i(t) - \mathbf{P}(\hat{a}_i)B^*y(t),$$

we have

$$(5.10) \quad \begin{aligned} \tilde{u}^i &\in \text{the bounded subset of } L^2(0, T; V), \quad \text{a.s. } \tilde{\mathcal{P}}, \\ \frac{d\tilde{u}^i}{dt} &\in \text{the bounded subset of } L^2(0, T; V'), \quad \text{a.s. } \tilde{\mathcal{P}}, \end{aligned}$$

and

$$(5.11) \quad \tilde{u}^i \rightarrow \hat{u}(\cdot, \tilde{a}) - \mathbf{P}(\tilde{a})B^*y(\cdot) \quad \text{strongly in } C(0, T; H) \cap L^2(0, T; V), \quad \text{a.s. } \tilde{\mathcal{P}},$$

where $\hat{u}(\cdot, \tilde{a})$ is a solution of (5.1a) for $\hat{a}_i = \tilde{a}$.

Proof. By using (5.9), (5.1a) can be represented by

$$(5.12) \quad \left\langle \frac{d\tilde{u}^i(t)}{dt} + (A(\hat{a}_i) + \mathbf{P}(\hat{a}_i)B^*B)(\tilde{u}^i(t) + \mathbf{P}(\hat{a}_i)B^*y(t)), \phi \right\rangle = 0, \quad \forall \phi \in V, \\ \tilde{u}_i(0) = 0.$$

Noting that $\mathbf{P}(\hat{a}_i) \in \mathcal{L}_2(V'; H) \cap \mathcal{L}_2(H; V)$, and $y \in C(0, T; R^m)$ almost surely $\tilde{\mathcal{P}}$, we have

$$\langle A(\hat{a}_i) + \mathbf{P}(\hat{a}_i)B^*B \phi_1, \phi_2 \rangle \geq \alpha \|\phi_1\|_V^2 - C(\alpha) \|\phi_1\|^2, \quad \exists \alpha, C(\alpha) > 0$$

and

$$(A(\hat{a}_i) + \mathbf{P}(\hat{a}_i)B^*B)\mathbf{P}(\hat{a}_i)B^*y(\cdot) \in V', \quad \forall t \in [0, T], \quad \text{a.s. } \tilde{\mathcal{P}}.$$

From [12, p. 106], (5.10) can be derived.

Denoting that \tilde{u} is a solution of

$$(5.13) \quad \left\langle \frac{d\tilde{u}(t)}{dt} + (A(\tilde{a}) + \mathbf{P}(\tilde{a})B^*B)(\tilde{u}(t) + \mathbf{P}(\tilde{a})B^*y(t)), \phi \right\rangle = 0, \quad \forall \phi \in V, \\ \tilde{u}(0) = 0,$$

we have

$$(5.14) \quad \begin{aligned} &|\tilde{u}^i(t) - \tilde{u}(t)|^2 + 2 \int_0^t \langle (A(\hat{a}_i) + \mathbf{P}(\hat{a}_i)B^*B)(\tilde{u}^i(s) - \tilde{u}(s)), \tilde{u}^i(s) - \tilde{u}(s) \rangle ds \\ &= 2 \int_0^t \langle (A(\tilde{a}) - A(\hat{a}_i)) + (\mathbf{P}(\tilde{a}) - \mathbf{P}(\hat{a}_i))B^*B \tilde{u}(s), \tilde{u}^i(s) - \tilde{u}(s) \rangle ds \\ &+ 2 \int_0^t \langle \{(A(\tilde{a})\mathbf{P}(\tilde{a}) - A(\hat{a}_i)\mathbf{P}(\hat{a}_i)) \\ &+ (\mathbf{P}(\tilde{a})B^*B\mathbf{P}(\tilde{a}) - \mathbf{P}(\hat{a}_i)B^*B\mathbf{P}(\hat{a}_i))\}B^*y(s), \tilde{u}^i(s) - \tilde{u}(s) \rangle ds \\ &= I_1 + I_2, \end{aligned}$$

for example. By using (5.6), we have

$$(5.15) \quad \begin{aligned} I_1 &= 2 \int_0^t \left(\tilde{a} - \hat{a}_i, \sum_{k=1}^n \frac{\partial \tilde{u}^i(s)}{\partial x_k} \frac{\partial (\tilde{u}^i(s) - \tilde{u}(s))}{\partial x_k} \right) ds \\ &+ 2 \int_0^t \langle (\mathbf{P}(\tilde{a}) - \mathbf{P}(\hat{a}_i))B^*B\tilde{u}(s), \tilde{u}^i(s) - \tilde{u}(s) \rangle ds \\ &\leq 2|\tilde{a} - \hat{a}_i| \left\{ \int_0^t \|\tilde{u}(s)\|_V^2 ds \int_0^t \|\tilde{u}^i(s) - \tilde{u}(s)\|_V^2 ds \right\}^{1/2} \\ &+ C|\tilde{a} - \hat{a}_i| \left\{ \int_0^t |\tilde{u}(s)|^2 ds \int_0^t |\tilde{u}^i(s) - \tilde{u}(s)|^2 ds \right\}^{1/2} \quad (\forall \varepsilon, \exists C(\varepsilon) > 0) \\ &\leq C(\varepsilon)|\tilde{a} - \hat{a}_i|^2 \int_0^T \|\tilde{u}(s)\|_V^2 ds + \varepsilon \int_0^T \|\tilde{u}^i(s) - \tilde{u}(s)\|_V^2 ds. \end{aligned}$$

Noting that \tilde{u} is also in the bounded set stated in (5.10), we find that there exists $\tilde{C}_1(\varepsilon) > 0$, for all $\varepsilon > 0$, such that

$$(5.16) \quad I_1 \leq \varepsilon \int_0^T \|\tilde{u}^i(s) - \tilde{u}(s)\|_V^2 ds + \tilde{C}_1(\varepsilon) |\tilde{a} - \hat{a}_i|^2.$$

Furthermore, from the fact that $B^*By(t) \in V$, for all $t \in [0, T]$, by using the same approach for (5.16), we get that there exists $\tilde{C}_2(\varepsilon) > 0$, for all $\varepsilon > 0$, such that

$$(5.17) \quad I_2 \leq \varepsilon \int_0^T \|\tilde{u}^i(s) - \tilde{u}(s)\|_V^2 ds + \tilde{C}_2(\varepsilon) |\tilde{a} - \hat{a}_i|^2 \sup_{0 \leq t \leq T} |y(t)|^2.$$

Hence, using Gronwall's inequality, we have

$$(5.18a) \quad \sup_{0 \leq t \leq T} |\tilde{u}^i(t) - \tilde{u}(t)|^2 \leq \text{const} |\tilde{a} - \hat{a}_i|^2$$

and

$$(5.18b) \quad \int_0^T \|\tilde{u}^i(t) - \tilde{u}(t)\|_V^2 dt \leq \text{const} |\tilde{a} - \hat{a}_i|^2.$$

From (5.18), (5.11) can be derived.

LEMMA 5.3. *Let \hat{a}_i be the same sequence in Lemma 5.1. Then we have*

$$(5.19) \quad \tilde{p}^i \rightarrow \tilde{p}(\cdot, \tilde{a}) \quad \text{strongly in } C(0, T; H) \cap L^2(0, T; V),$$

where $\tilde{p}(\cdot, \tilde{a})$ is a solution of (5.2a) for $\hat{a}_i = \tilde{a}$ and

$$(5.20) \quad \mathbf{R}(\hat{a}_i)B^* \rightarrow \mathbf{R}(\tilde{a})B^* \quad \text{strongly in } \mathcal{L}(R^m; H).$$

The proof of this lemma is omitted, since by applying the same procedure used in the proof of the previous lemmas, this proof can easily be obtained.

LEMMA 5.4. *Letting \hat{a}_i be in $\Theta_R \cap \{a \mid \|a\|_{H^s} \leq \text{const}\}$, we find that the solution \hat{a}_{i+1} of the variational inequality (5.4) satisfies the following condition:*

$$(5.21) \quad \|\hat{a}_{i+1}\|_{H^s} \leq C,$$

where the constant C is independent of \hat{a}_i .

Proof. From the results of Lemmas 5.2 and 5.3, for $\hat{a}_i \in$ the bounded subset of Θ_R , we have

$$(5.22a) \quad \left\| \int_0^T \sum_{k=1}^n \frac{\partial \hat{u}^i}{\partial x_k} \frac{\partial \tilde{p}^i}{\partial x_k} dt \right\|_{(H^s)'} \leq C \left(\int_0^T \|\hat{u}^i\|_V^2 dt + \int_0^T \|\tilde{p}^i(t)\|_V^2 dt \right) \leq C$$

and

$$(5.22b) \quad \left\| \sum_{k=1}^n \int_G \frac{\partial R(\hat{a}_i, x, z)}{\partial x_k} \frac{\partial P(\hat{a}_i, z, x)}{\partial x_k} dz \right\|_{(H^s)'}^2 \leq C,$$

where C is a constant independent of \hat{a}_i . From Theorem 2.1 in Kinderlehrer [9, p. 24], we find that the variational inequality (5.4) has a unique solution $\hat{a}_{i+1} \in \Theta_R$. Setting $a = \alpha$ in (5.4), we have

$$2\gamma(T) \|\hat{a}_{i+1}\|_{H^s}^2 \leq -(f(\hat{a}_i), \alpha - \hat{a}_{i+1}) + 2\gamma(T)(\hat{a}_{i+1}, \alpha)$$

(from (5.22))

$$\leq \varepsilon \|\hat{a}_{i+1}\|_{H^s}^2 + \text{const independent of } \hat{a}_{i+1}, \quad \forall \varepsilon > 0.$$

The estimate (5.21) is derived.

Proof of Theorem 5.1. From the results of Lemmas 5.1–5.4, noting that \hat{a}_0 is in the bounded subset of Θ_R , the whole sequence \hat{a}_i satisfies (5.21). Then we can extract a subsequence, still denoted by \hat{a}_i such that

$$(5.23) \quad \hat{a}_i \rightarrow \tilde{a} \quad \text{weakly in } \Theta_R \text{ and strongly in } H.$$

Hence, the remaining task is to show that \tilde{a} is a solution of the variational inequality (4.33). From Lemmas 5.1–5.3, we have

$$(5.24) \quad f(\hat{a}_i) \rightarrow f(\tilde{a}) \quad \text{strongly in } (H^s)' \quad \text{a.s. } \tilde{\mathcal{P}}.$$

Noting that

$$(5.25) \quad \langle \Lambda \hat{a}_{i+1}, a \rangle \rightarrow \langle \Lambda \tilde{a}, a \rangle$$

and

$$-\liminf \langle \Lambda \hat{a}_{i+1}, \hat{a}_{i+1} \rangle \leq -\langle \Lambda \tilde{a}, \tilde{a} \rangle,$$

we can take a limit in (5.4) as $i \rightarrow \infty$ to derive the variational inequality (4.33).

6. Numerical examples. In § 2 we assume that the system input is only a noise term. However, by applying the deterministic term $g(x) \in L^2(G)$ to the system equation (2.1), we can obtain the same results as those in the previous sections. To illustrate the theory, we will consider the following simple system with the given deterministic input $g(x)$:

$$(6.1a) \quad u(t, x) - \int_0^t \frac{\partial}{\partial x} \left(a^0(x) \frac{\partial u(s, x)}{\partial x} \right) ds = u_0(x) + tg(x) + w(t, x) \quad \text{in } [0, t_f] \times [0, 1],$$

$$(6.1b) \quad u(t, 0) = u(t, 1) = 0 \quad \text{on } [0, t_f],$$

$$(6.2) \quad y_i(t) = \int_0^t \int_{G_i} b_i(x) u(s, x) dx ds + v_i(t)$$

for $i = 1, 2, \dots, m$. We use the well-known formula of finite-difference scheme to perform the digital simulation experiments (see, e.g., [7]). Hence, we set $m = 12$, $\Delta x = 1/26$, $G_i =]2i - 1, 2i + 1[$, $i = 1, 2, \dots, 12$. The deterministic input $g(x)$ and the weight function $b_i(x)$ are given by

$$g(x)_{x=j\Delta x} = \begin{cases} 0 & j = 1, 2, \dots, 5 \\ 100 & j = 6 \\ 0 & j = 7, 8, \dots, 12 \\ -100 & j = 13 \\ 0 & j = 14, \dots, 18 \\ 100 & j = 19 \\ 0 & j = 20, \dots, 25 \end{cases}$$

and

$$b_i(x)_{x=j\Delta x} = \begin{cases} b\Delta x & j = 2i - 1 \\ 2b\Delta x & j = 2i \\ b\Delta x & j = 2i + 1 \\ 0 & j = \text{others} \end{cases},$$

respectively. The system and observation noises are generated as follows:

$$\frac{\partial w(t, x)}{\partial t} \cong N_{i,j}/\sqrt{\Delta t \Delta x} \quad \text{and} \quad \frac{dv(t)}{dt} \cong N_k/\sqrt{\Delta t},$$

where $N_{i,j}$ and N_k are mutually independent Gaussian random numbers, i.e., $N_{i,j} \in N(0, \sigma_s)$ and $N_k \in N(0, 1)$. The incremental covariance Q is given by

$$Q = \int_0^1 q(x, z)(\cdot) dz \cong \sum_{k=1}^{26} \delta_{i,k}(\cdot) \quad \text{for } x \cong i\Delta x \quad \text{and} \quad z \cong k\Delta x (1 \leq i, k \leq 26).$$

Equations (6.1) and (6.2) are transformed into the finite-difference equations by using the Crank–Nicolson scheme. Furthermore, to realize the stationary system state in the digital simulation experiments, the initial condition is set as

$$u_0(x) = \left(-\frac{\partial}{\partial x} (a^0(x)) \frac{\partial}{\partial x} \right)^{-1} g(x) + N_I(x),$$

where N_I is a Gaussian noise, i.e., $N_I(j\Delta x) \in N(0, \sigma_s)$.

Setting the regularization operator Λ as

$$\langle\langle \Lambda \phi, \phi \rangle\rangle = \left(\frac{d^3 \phi}{dx^3}, \frac{d^3 \phi}{dx^3} \right) + |\phi|^2,$$

we consider the four cases given in Table 1. In these numerical examples, we should add the extra terms $E\{u_0(x)\}$ and $tg(x)$ to the right-hand side of (5.1a). Here we formally let $E\{u_0(x)\}$ be

$$\left(-\frac{\partial}{\partial x} (a^0(x)) \frac{\partial}{\partial x} \right)^{-1} g(x),$$

as the numerical data.

To carry out these simulation experiments, all equations were approximated to the finite-difference equations used in [8] and [11] and solved by a digital computer. In Figs. 1–4, corresponding to the true parameters given in Table 1, sample runs of the estimates generated by the proposed algorithm are demonstrated. We can perform the algorithm without a regularization term because all equations of the algorithm are approximated to the finite-difference equations. (Of course, we cannot prove the convergence property of the estimate without a regularization term.) Comparing the sample runs of the RMLE and the MLE as shown in Figs. 1 and 2, respectively, we can safely conclude that the regularization method works well. In this paper, we cannot show that the uniqueness of the solution to the nonlinear variational inequality (4.33) as well as we have shown the exact matching of the parameter estimate and the true parameter. However, the results of the numerical examples in Figs. 3 and 4 seem to show the possibility of proving the identifiability of the regularized estimate. From

TABLE 1
Numerical values for identification of $a^0(x)$.

Case	True $a^0(x)$	σ_s	b	γ
1	$1 + \sin^4(\pi x)$	1×10^{-2}	10^6	10^{-8}
2	$1 + \sin^4(\pi x)$	1×10^{-2}	10^6	0
3	$1 + x$	2×10^{-1}	10^4	10^{-8}
4	$1 + \sin^2(2\pi x)$	2×10^{-1}	10^4	10^{-8}

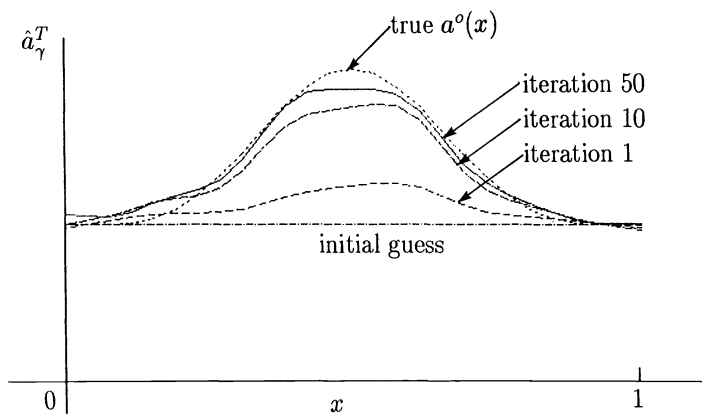


FIG. 1. Sample runs of RMLE for Case 1.

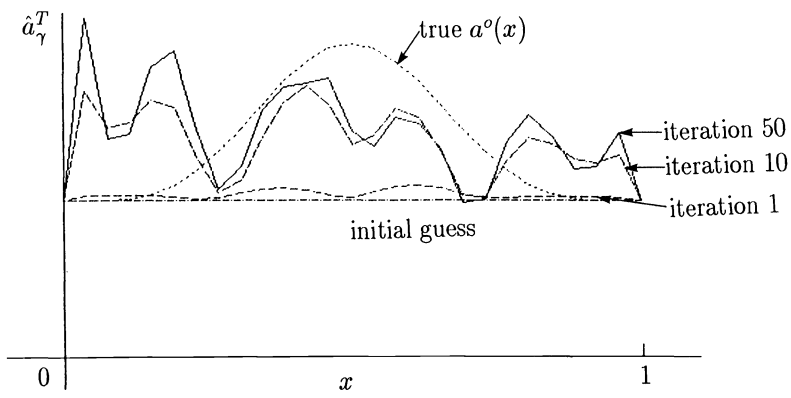


FIG. 2. Sample runs of MLE for Case 2.

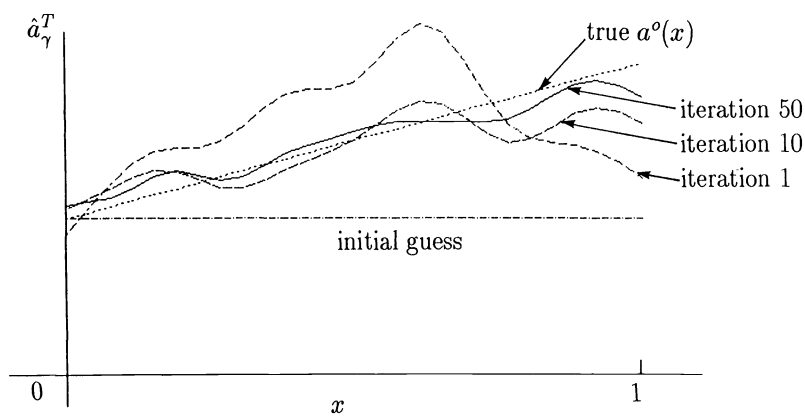


FIG. 3. Sample runs of RMLE for Case 3.

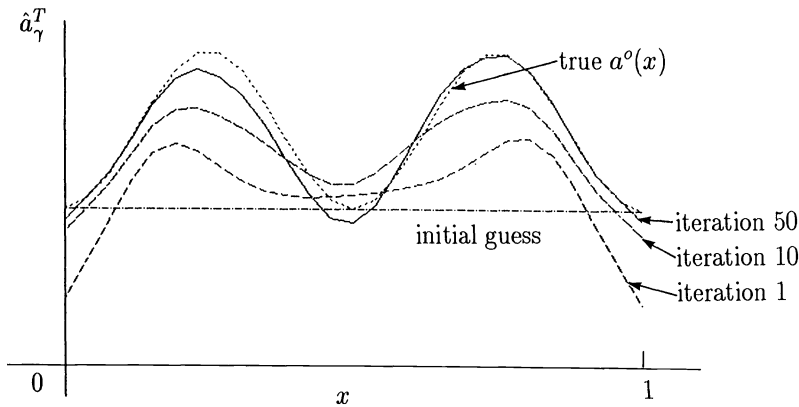


FIG. 4. Sample runs of RMLE for Case 4.

results of simulation experiments, we can conclude that the deterministic input $g(x)$ and the initial condition $E\{u_0(x)\}$ strongly depend on the parameter estimate.

Acknowledgments. The author thanks Tasuku Hoshino, who carried out the numerical calculation in this paper. Thanks are extended to the anonymous referees for making several suggestions and comments that enhanced the presentation of the report.

REFERENCES

- [1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
- [2] S. I. AIHARA AND A. BAGCHI, *Parameter identification for stochastic diffusion equations with unknown boundary conditions*, *Appl. Math. Optim.*, 17 (1988), pp. 15–36.
- [3] S. I. AIHARA AND Y. SUNAHARA, *Identification of an infinite-dimensional parameter for stochastic diffusion equations*, *SIAM J. Control Optim.*, 26 (1988), pp. 1062–1075.
- [4] A. BAGCHI AND V. BORKAR, *Parameter identification in infinite-dimensional linear systems*, *Stochastics*, 12 (1984), pp. 472–486.
- [5] V. E. BENES AND I. KARATZAS, *On the relation of Zakai's and Mortensen's equations*, *SIAM J. Control Optim.*, 21 (1983), pp. 472–489.
- [6] A. BENSOUSSAN, *Filtrage Optimal des Systems Lineaires*, Dunod, Paris, 1971.
- [7] G. CHAVENT AND P. LEMONIER, *Identification de la non-linearité d'une équation parabolique quasilineaire*, *Appl. Math. Optim.*, 1 (1974), pp. 121–162.
- [8] G. GLOWINSKI, J. L. LIONS, AND R. TREMOLIERES, *Analyse Numérique des Inéquations Variationnelles*, Dunod, Paris, 1976.
- [9] D. KINDERLEHRER AND G. STAMPACCHIA, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980.
- [10] P. KOTELENEZ AND R. F. CURTAIN, *Local behavior of Hilbert space valued stochastic integrals and the continuity of mild solutions of stochastic evolution equations*, *Stochastics*, 6 (1982), pp. 239–257.
- [11] C. KRAVARIS AND J. H. SEINFELD, *Identification of parameters in distributed parameter systems by regularization*, *SIAM J. Control Optim.*, 23 (1985), pp. 217–241.
- [12] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, New York, 1971.
- [13] R. S. LIPTSER AND A. N. SHIRYAEV, *Statistics of Random Processes I*, Springer-Verlag, Berlin, New York, 1971.
- [14] S. MIZOHATA, *The Theory of Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1973.
- [15] J. Y. OUVARD, *Martingale projection and linear filtering in Hilbert spaces I: Theory*, *SIAM J. Control Optim.*, 16 (1978), pp. 912–937.

- [16] E. PARDOUX, *Équation aux dérivées partielles stochastiques non linéaires monotones*, Ph.D. thesis, Université Paris XI, 1975.
- [17] J. K. TUGNAIT, *Continuous-time system identification in compact parameter sets*, IEEE Trans. Inform. Theory, IT-31 (1981), pp. 652–659.
- [18] J. ZABCZYK, *Remarks on the algebraic Riccati equation in Hilbert spaces*, Appl. Math. Optim., 3 (1976), pp. 251–259.

ALGEBRAIC RICCATI EQUATIONS AND THE DISTANCE TO THE NEAREST UNCONTROLLABLE PAIR*

P. GAHINET†‡ AND A. J. LAUB†

Abstract. A connection is established between nearness to unstabilizability of a stabilizable pair (A, B) of matrices and nearness to singularity of the symmetric positive definite solution to an associated algebraic Riccati equation. From this result, computable upper and lower bounds are derived for the distance of (A, B) to the nearest uncontrollable pair. Numerical tests confirm the validity of the method and potential applications are discussed.

Key words. Riccati equation, nearness to uncontrollability, stabilizability, robustness

AMS(MOS) subject classifications. 49E30, 93B35, 93B40

1. Introduction. When numerically assessing whether a pair of matrices $(A, B) \in \mathbf{R}^{n \times n} \times \mathbf{R}^{n \times r}$ is controllable (or stabilizable), tests that simply provide a yes/no answer are not entirely satisfactory [17], [18]. Instead, an estimate of how far the pair is from the set of uncontrollable (respectively, unstabilizable) pairs is more relevant. Unfortunately, this involves a nonconvex minimization in a space of n dimensions, and existing numerical methods to search for minima often suffer from the following limitations: the computed minima are only local; a two-dimensional search is necessary when complex perturbations are allowed; and the speed of convergence is guaranteed to be quadratic only in the proximity of the local minima, and a high computational overhead may thus be attached.

Few lower or upper bounds on the distance to uncontrollability are available in the literature. Upper bounds were proposed in [1] but they require either forming the controllability matrix or that A be stable. A lower bound was obtained by Demmel in [6]. Finally, an expression for the distance to unstabilizability was obtained in [13, Prop. 4.8] in the special case where only A is perturbed. Thus existing bounds often have restrictive conditions of validity. Moreover, they can be very conservative in some cases.

In this paper, nearness to uncontrollability of a controllable pair (A, B) is related to nearness to singularity of the positive definite solution to an associated algebraic Riccati equation (ARE). This connection provides lower and upper bounds that are relatively tight in most cases, computable at a reasonable cost, and have a simple interpretation in system theoretical terms. This approach applies to the most general pair (A, B) and entirely departs from the usual formulation as a functional minimization problem.

The paper is organized as follows. First, the concept of nearness to uncontrollability and its formulation as a nonconvex minimization problem are reviewed (§ 2). Classical numerical methods to find local minima are also recalled in § 3. A new result concerning the distance to uncontrollability is then presented in § 4. In the second part, it is shown how stabilizability robustness can be assessed via an ARE (§ 5), and lower and upper

* Received by the editors December 6, 1989; accepted for publication (in revised form) March 27, 1991. This research was supported by National Science Foundation (and Air Force Office of Scientific Research) grant ECS87-18897 and Air Force Office of Scientific Research contract AFOSR-91-0240.

† Department of Electrical and Computer Engineering, University of California, Santa Barbara, California 93106.

‡ Present address, Institut National de Recherche en Informatique et en Automatique, Domaine de Voluceau, BP 105, 78153 Le Chesnay Cedex, France.

bounds are derived from this connection (§§ 7, 8). Finally, the results of extensive numerical testing are analyzed, and potential applications of this new tool are discussed.

2. Definitions, notation, and elementary remarks. Consider the linear dynamic system

$$(2.1) \quad \dot{x} = Ax + Bu,$$

where $x \in \mathbf{R}^n$ denotes the state vector, $u \in \mathbf{R}^r$ the input or control vector, and $(A, B) \in \mathbf{R}^{n \times n} \times \mathbf{R}^{n \times r}$ with $r \leq n$. The pair (A, B) is controllable if and only if the pencil $(A - sI, B)$ has full row rank n for any complex s . Similarly, (A, B) is stabilizable if and only if $(A - sI, B)$ has rank n for any complex s with nonnegative real part; in other words, if and only if (A, B) has no uncontrollable mode in the closed right half-plane.

The distance of (A, B) to the nearest unstabilizable or uncontrollable pair is defined as the norm of the smallest perturbation $(\delta A, \delta B)$, which makes the pair $(A + \delta A, B + \delta B)$ unstabilizable or uncontrollable, respectively. Throughout the paper, the Frobenius norm $\|M\|_F = (\text{Trace}(MM^T))^{1/2}$ will be used for measuring the perturbation magnitude. The perturbations considered can be either complex or restricted to the real field, which leads to the definition of two distinct distance measures. The distance of (A, B) to the nearest unstabilizable pair will be denoted by $\nu_{\mathbf{R}}$ when only real perturbations are considered, and $\nu_{\mathbf{C}}$ when complex perturbations are allowed. That is,

$$(2.2) \quad \nu_{\mathbf{F}}(A, B) = \inf \{(\|\delta A\|_F^2 + \|\delta B\|_F^2)^{1/2}; \delta A \in \mathbf{F}^{n \times n}; \delta B \in \mathbf{F}^{n \times r}; \\ (A + \delta A, B + \delta B) \text{ unstabilizable}\},$$

where $\mathbf{F} = \mathbf{R}, \mathbf{C}$.

A parallel definition is given below to the distance of (A, B) to the nearest uncontrollable pair, which will be referred to as $\mu_{\mathbf{R}}$ (real perturbations) and $\mu_{\mathbf{C}}$ (complex perturbations):

$$(2.3) \quad \mu_{\mathbf{F}}(A, B) = \inf \{(\|\delta A\|_F^2 + \|\delta B\|_F^2)^{1/2}; \delta A \in \mathbf{F}^{n \times n}; \delta B \in \mathbf{F}^{n \times r}; \\ (A + \delta A, B + \delta B) \text{ uncontrollable}\}.$$

Under a state coordinate transformation $x' = Tx$ in (2.1), the pair (A, B) becomes (TAT^{-1}, TB) . Note that $\nu_{\mathbf{R}}$ and $\mu_{\mathbf{R}}$ (respectively, $\nu_{\mathbf{C}}$ and $\mu_{\mathbf{C}}$) are invariant for T orthogonal (respectively, unitary) but are generally affected by other transformations T [9], [20].

Note finally that controllability and stabilizability are generic properties; that is, the set of controllable (or stabilizable) pairs is open and dense. Consequently, $\mu(A, B) = 0$ if and only if (A, B) is uncontrollable (and similarly for ν). Also, the infimum in (2.2) or (2.3) is attained for some perturbation. We will refer to such a perturbation as ν -minimal or μ -minimal, respectively.

The distances $\mu_{\mathbf{C}}$ and $\nu_{\mathbf{C}}$ can be related to nearness to rank deficiency as shown in [10]. Specifically,

$$(2.4) \quad \nu_{\mathbf{C}} = \min_{\text{Re } \lambda \geq 0} \sigma_{\min}(A - \lambda I, B),$$

$$(2.5) \quad \mu_{\mathbf{C}} = \min_{\lambda \in \mathbf{C}} \sigma_{\min}(A - \lambda I, B),$$

where $\sigma_{\min}(\cdot)$ denotes the minimum singular value.

The minimization (2.5) can be reformulated as a minimization over the unit sphere in \mathbf{C}^n , that is,

$$(2.6) \quad \mu_{\mathbf{C}}(A, B) = \min_{q \in \mathbf{C}^n, \|q\|=1} F_{\mathbf{C}}(q)^{1/2}, \quad \text{where } F_{\mathbf{C}}(q) = q^H(AA^T + BB^T)q - |q^H Aq|^2.$$

The two minimization problems are dual and their local extrema are any pair $(\lambda^*, q^*) \in \mathbb{C} \times \mathbb{C}^n$ satisfying the following conditions:

- (C1) λ^* is extremal for (2.5),
- (C2) q^* is the left singular vector associated with the smallest singular value of $(A - \lambda^* I, B)$,
- (C3) $\lambda^* = q^{*H} A q^*$.

These duality results can be found in [2] and [22] and play a crucial role in the design of descent algorithms (see § 3).

A consequence of this characterization is the following result.

PROPOSITION 2.1. *Let $(\delta A, \delta B)$ be a $\mu_{\mathbb{C}}$ -minimal perturbation of (A, B) . Then $(\delta A, \delta B)$ is rank-one, and there exists a unitary state transformation U such that*

$$(2.7) \quad \begin{aligned} \delta A &= U \begin{pmatrix} O_{n-1} & 0 \\ -\delta a^H & 0 \end{pmatrix} U^H, & \delta B &= U \begin{pmatrix} 0 \\ -\delta b^H \end{pmatrix}, \\ \delta a &\in \mathbb{C}^{(n-1) \times 1}, & \delta b &\in \mathbb{C}^{r \times 1}, \quad \|\delta a\|^2 + \|\delta b\|^2 = \mu_{\mathbb{C}}^2. \end{aligned}$$

Moreover, the representation of (A, B) with respect to the new basis is given by

$$(2.8) \quad U^H A U = \begin{pmatrix} A_{11} & a_{12} \\ \delta a^H & \lambda^* \end{pmatrix}, \quad U^H B = \begin{pmatrix} B_1 \\ \delta b^H \end{pmatrix},$$

where $\lambda^* \in \mathbb{C}$ is minimal in (2.5).

Proof. See Appendix A.

When restricting ourselves to real perturbations, the minimizing perturbation in (2.5) may be either rank-one or rank-two. Specifically, $\mu_{\mathbb{R}}$ -minimal perturbations can be categorized as follows.

PROPOSITION 2.2. *Let $(\delta A, \delta B)$ be a $\mu_{\mathbb{R}}$ -minimal perturbation of (A, B) . Then $(\delta A, \delta B)$ is either rank-one or rank-two, and there exists an orthogonal state transformation U such that either*

$$(2.9) \quad \begin{aligned} \delta A &= U \begin{pmatrix} O_{n-1} & 0 \\ -\delta a^T & 0 \end{pmatrix} U^T, & \delta B &= U \begin{pmatrix} 0 \\ -\delta b^T \end{pmatrix}, \\ \delta a &\in \mathbb{R}^{(n-1) \times 1}, & \delta b &\in \mathbb{R}^{r \times 1}, \quad \|\delta a\|^2 + \|\delta b\|^2 = \mu_{\mathbb{R}}^2, \end{aligned}$$

and the representation of (A, B) with respect to the new basis is

$$(2.10) \quad U^T A U = \begin{pmatrix} A_{11} & a_{12} \\ \delta a^T & \lambda^* \end{pmatrix}, \quad U^T B = \begin{pmatrix} B_1 \\ \delta b^T \end{pmatrix},$$

where $\lambda^* \in \mathbb{R}$ minimizes $\sigma_{\min}(A - \lambda I, B)$ over all real λ , or

$$(2.11) \quad \begin{aligned} \delta A &= U \begin{pmatrix} O_{n-2} & 0 \\ -\delta A_{21}^T & 0 \end{pmatrix} U^T, & \delta B &= U \begin{pmatrix} 0 \\ -\delta B_2^T \end{pmatrix}, \\ \delta A_{21} &\in \mathbb{R}^{(n-2) \times 2}, & \delta B_2 &\in \mathbb{R}^{r \times 2}, \quad \|\delta A_{21}\|_F^2 + \|\delta B_2\|_F^2 = \mu_{\mathbb{R}}^2, \end{aligned}$$

and the representation of (A, B) with respect to the new basis is

$$(2.12) \quad U^T A U = \begin{pmatrix} A_{11} & A_{12} \\ \delta A_{21}^T & A_{22} \end{pmatrix}, \quad U^T B = \begin{pmatrix} B_1 \\ \delta B_2^T \end{pmatrix},$$

where $A_{22} \in \mathbb{R}^{2 \times 2}$ has complex conjugate eigenvalues with nonzero imaginary parts.

Proof. See Appendix A.

The perturbations of type (2.9) will be called one-dimensional, and those of type (2.11), two-dimensional. This distinction leads to the introduction of two separate

distance measures, denoted by $\mu_{\mathbf{R},1}$ and $\mu_{\mathbf{R},2}$, which correspond to the norm of a minimal real one-dimensional and two-dimensional perturbation, respectively. Note that

$$(2.13) \quad \mu_{\mathbf{R}}(A, B) = \min(\mu_{\mathbf{R},1}(A, B), \mu_{\mathbf{R},2}(A, B)),$$

$$\mu_{\mathbf{R},1}(A, B) = \min_{\lambda \in \mathbf{R}} \delta_{\min}(A - \lambda I, B),$$

$$(2.14) \quad \mu_{\mathbf{R},1}(A, B) = \min_{q \in \mathbf{R}^n, \|q\|=1} F_{\mathbf{R},1}(q)^{1/2},$$

$$\text{where } F_{\mathbf{R},1}(q) = q^T(AA^T + BB^T)q - (q^T Aa)^2.$$

Moreover, a local extremum (λ^*, q^*) of (2.13) is characterized by

(R1) λ^* is extremal for (2.13),

(R2) q^* is the left singular vector associated with the smallest singular value of $(A - \lambda^* I, B)$,

(R3) $\lambda^* = q^{*T} A q^*$.

Finally, the analogue of (2.14) for two-dimensional real perturbations is [22]

$$(2.15) \quad \mu_{\mathbf{R},2}(A, B) = \min_{Q \in \mathbf{R}^{n \times 2}, Q^T Q = I} F_{\mathbf{R},2}(Q)^{1/2},$$

where

$$F_{\mathbf{R},2}(Q) = \text{Trace} \{ Q^T (AA^T + BB^T) Q - Q^T A Q Q^T A^T Q \}.$$

To conclude this section, note the simple connection between distance to uncontrollability and distance to unstabilizability.

PROPOSITION 2.3. *The measures $\mu_{\mathbf{F}}$ and $\nu_{\mathbf{F}}$, with $\mathbf{F} = \mathbf{R}, \mathbf{C}$, are related by*

$$(2.16) \quad \mu_{\mathbf{F}}(A, B) = \min(\nu_{\mathbf{F}}(A, B), \nu_{\mathbf{F}}(-A, B)).$$

Proof. A pair (A, B) is controllable if and only if both (A, B) and $(-A, B)$ are stabilizable. Equivalently, a pair (A, B) is uncontrollable if and only if either (A, B) or $(-A, B)$ is unstabilizable. Consequently, a perturbation $(\delta A, \delta B)$ renders (A, B) uncontrollable if and only if $(\delta A, \delta B)$ makes (A, B) unstabilizable or $(-\delta A, \delta B)$ makes $(-A, B)$ unstabilizable. Identity (2.16) immediately follows from this remark and definitions (2.2) and (2.3).

3. Algorithms to estimate $\mu_{\mathbf{R}}$ and $\mu_{\mathbf{C}}$: A brief survey. Most existing algorithms use descent methods to find local minima of (2.5), (2.6), (2.13), (2.14), or (2.15). The various methods can be classified in two types: (1) descent schemes that search for a complex (respectively, real) λ , which locally minimizes (2.5) or (2.13), and (2) descent methods that compute a sequence of elementary rotations [20] to find local minima of $F_{\mathbf{R},1}$, $F_{\mathbf{R},2}$ or $F_{\mathbf{C}}$ as defined in (2.14), (2.15), and (2.6), respectively.

The use of descent algorithms to find minimizing λ 's in (2.5) or (2.6) originated in [8]. Such algorithms can only estimate $\mu_{\mathbf{C}}$ or $\mu_{\mathbf{R},1}$. Their speed of convergence is at least quadratic in the neighborhood of local minima (see [2]), but their main disadvantage is that they provide only local minima. Finding all such minima in the one-dimensional real case is relatively easy. In the complex case, however, their retrieval requires a recursive subdivision and scanning of the complex plane (see [4]), which may imply a high computational overhead.

The second type of algorithm aims at decreasing the functionals $F_{\mathbf{R},1}$, $F_{\mathbf{R},2}$, or $F_{\mathbf{C}}$. A thorough description of such schemes can be found in [22]. This approach has the advantage of being general, but suffers the same difficulties with handling local

minima. In fact, it is only guaranteed to converge to a local minimum for one of the singular values of $(A - \lambda I, B)$. Moreover, it is unclear how to methodically retrieve all local minima.

4. The relation between μ_C and μ_R . The computation of μ_R has traditionally been considered more difficult than that of μ_C , mainly because $\mu_{R,2}$ cannot be obtained through a one-dimensional minimization as in (2.5) or (2.13). In this section, we present a new result that bounds μ_R from above and below in terms of μ_C . For μ ranging between 10^{-6} and 1, a typical range of interest in applications, these bounds indicate that μ_R or μ_C are of approximately the same order of magnitude. Thus, it is sufficient to estimate μ_C even when only real perturbations are of interest. Also, the lower bound on μ_C obtained in § 7 and the upper bound on μ_R derived in § 8 can be expected to be reasonable lower and upper estimates for both μ_R and μ_C .

The following technical lemma is needed.

LEMMA 4.1. *Let $q \in \mathbb{C}^n$ be a unit vector minimizing F_C as defined in (2.6). Then there exist two orthogonal real vectors u and v with $\|u\|^2 + \|v\|^2 = 1$, such that $w = u + iv$ and $\bar{w} = u - iv$ minimize F_C as well.*

Proof. See [21].

The next theorem bounds the gap between μ_R and μ_C .

THEOREM 4.2. *The distance measures $\mu_R(A, B)$ and $\mu_C(A, B)$ are related by*

$$(4.1) \quad \mu_C \leq \mu_R \leq \max(2\sqrt{2} \mu_C, 2\|A - A^T\|^{1/3} \mu_C^{2/3}).$$

Proof. The first inequality is trivial.

Let $q \in \mathbb{C}^n$ be such that $F_C(q) = \mu_C^2(A, B)$. Without loss of generality, $q = u + iv$ can be chosen so that the real vectors u and v are orthogonal, from Lemma 4.1. Since q is an extremum of F_C , the characterization (C1)–(C3) of § 1 ensures the existence of a unit vector t such that

$$(4.2) \quad q^H(A - (q^H A q)I, B) = \mu_C t^H.$$

Write $t^H = (r^H, s^H)$ with $r \in \mathbb{C}^n$ and $s \in \mathbb{C}^r$ and define $\lambda = q^H A q$. Then (4.2) is equivalent to

$$q^H A = \lambda q^H + \mu_C r^H, \quad q^H B = \mu_C s^H.$$

Taking the real and imaginary parts of these last two equations yields

$$(4.3) \quad \begin{aligned} u^T A &= \operatorname{Re} \lambda u^T + \operatorname{Im} \lambda v^T + \mu_C \operatorname{Re} r^T, & u^T B &= \mu_C \operatorname{Re} s^T, \\ v^T A &= -\operatorname{Im} \lambda u^T + \operatorname{Re} \lambda v^T + \mu_C \operatorname{Im} r^T, & v^T B &= \mu_C \operatorname{Im} s^T. \end{aligned}$$

Suppose that $\|u\| \cdot \|v\| \neq 0$, and let $n_u = \|u\|$ and $n_v = \|v\|$. Choose $n-2$ vectors e_1, \dots, e_{n-2} such that the set $(e_1, \dots, e_{n-2}, u/n_u, v/n_v)$ is orthonormal. Then with respect to this basis the pair (A, B) has the form

$$A = \begin{pmatrix} * & * \\ A_{21} & A_{22} \end{pmatrix}, \quad B = \begin{pmatrix} * \\ B_2 \end{pmatrix},$$

where

$$A_{21} = \begin{pmatrix} \left(\frac{\mu_C}{n_u}\right) \operatorname{Re} r^T \\ \left(\frac{\mu_C}{n_v}\right) \operatorname{Im} r^T \end{pmatrix}, \quad A_{22} = \begin{pmatrix} \operatorname{Re} \lambda & \left(\frac{n_v}{n_u}\right) \operatorname{Im} \lambda \\ -\left(\frac{n_u}{n_v}\right) \operatorname{Im} \lambda & \operatorname{Re} \lambda \end{pmatrix}, \quad B_2 = \begin{pmatrix} \left(\frac{\mu_C}{n_u}\right) \operatorname{Re} s^T \\ \left(\frac{\mu_C}{n_v}\right) \operatorname{Im} s^T \end{pmatrix}.$$

It follows that

$$(4.4) \quad \begin{aligned} \mu_{\mathbf{R}} &\leq (\|A_{21}\|_F^2 + \|B_2\|_F^2)^{1/2} \leq \mu_C \max\left(\frac{1}{n_u}, \frac{1}{n_v}\right) (\|r\|^2 + \|s\|^2)^{1/2} \\ &\leq \frac{\mu_C}{\min(n_u, n_v)}, \end{aligned}$$

using the fact that $\|r\|^2 + \|s\|^2 = \|t\|^2 = 1$.

The right-hand side of (4.4) becomes unbounded when either $\|u\|$ or $\|v\|$ approaches zero. In that case, however, q approaches a real vector and an alternate bound can be derived as follows. Returning to (4.3), rewrite the top two equations as

$$u^T(A - \operatorname{Re} \lambda I, B) = \mu_C \operatorname{Re} t^T + \operatorname{Im} \lambda (v^T, 0).$$

Since $\mu_{\mathbf{R},1} \leq \|(u^T/n_u)(A - \operatorname{Re} \lambda I, B)\|$, it follows that

$$\mu_{\mathbf{R}} \leq \mu_{\mathbf{R},1} \leq \frac{\mu_C + |\operatorname{Im} \lambda| n_v}{n_u}.$$

Now,

$$|\operatorname{Im} \lambda| = |u^T A v - v^T A u| = |u^T (A - A^T) v| \leq \|A - A^T\| n_u n_v.$$

Therefore,

$$(4.5) \quad \mu_{\mathbf{R}} \leq \frac{\mu_C}{n_u} + \|A - A^T\| n_v^2 \quad \text{with} \quad n_u^2 + n_v^2 = 1.$$

A similar manipulation starting with the bottom two equations in (4.3) leads to the counterpart (4.5') of (4.5), with u and v interchanged. Let $x := \min(n_u, n_v)$ and $\alpha := \|A - A^T\|$. Combining (4.5) and (4.5') yields the following upper bound for $\mu_{\mathbf{R}}$:

$$(4.6) \quad \mu_{\mathbf{R}} \leq \frac{\mu_C}{\sqrt{1-x^2}} + \alpha x^2 \leq \sqrt{2} \mu_C + \alpha x^2,$$

where the second inequality is obtained upon noting that x ranges in the interval $J = [0, \sqrt{2}/2]$.

Combining (4.4) and (4.6) yields $\mu_{\mathbf{R}} \leq \min(f(x), g(x))$, where the functions f and g are defined as

$$f(x) = \frac{\mu_C}{x}, \quad g(x) = \sqrt{2} \mu_C + \alpha x^2, \quad x = \min(\|u\|, \|v\|).$$

Since $x \in J$, an upper bound for $\mu_{\mathbf{R}}$ is obtained as $\xi = \max_{x \in J} \min(f(x), g(x))$. Now, the functions f and g are monotonically decreasing and increasing over J , respectively. Moreover, $f(x) \rightarrow +\infty$ as $x \rightarrow 0$, and $f(\sqrt{2}/2) = \sqrt{2} \mu_C$; $g(\sqrt{2}/2) = \sqrt{2} \mu_C + \alpha/2$. Therefore, there is a unique point x_0 in J such that $f(x_0) = g(x_0)$ and ξ can be expressed as

$$(4.7) \quad \xi = f(x_0) = \min_{x \in J} \max(f(x), g(x)).$$

We now conclude by bounding from above the rightmost term in (4.7). This is done by distinguishing between the following two cases.

(1) If $\mu_C \leq 2^{-3/2} \alpha$, then $x_1 = \alpha^{-1/3} \mu_C^{1/3}$ is in J and

$$f(x_1) = \alpha^{1/3} \mu_C^{2/3}, \quad g(x_1) = \sqrt{2} \mu_C + \alpha^{1/3} \mu_C^{2/3} = \alpha^{1/3} \mu_C^{2/3} (1 + \sqrt{2} x_1) \leq 2 \alpha^{1/3} \mu_C^{2/3}.$$

From (4.7), it follows that $\xi \leq 2 \alpha^{1/3} \mu_C^{2/3}$, and (4.1) is obtained upon noting that $2 \alpha^{1/3} \mu_C^{2/3} \geq 2 \sqrt{2} \mu_C$ when $\mu_C \leq 2^{-3/2} \alpha$.

- (2) If $\mu_C > 2^{-3/2}\alpha$, then $2\alpha^{1/3}\mu_C^{2/3} < 2\sqrt{2}\mu_C$. Also, $g(x) \leq \sqrt{2}\mu_C + \alpha/2 < 2\sqrt{2}\mu_C$, whence $\xi < 2\sqrt{2}\mu_C$ and (4.1) holds again.

Therefore, the upper bound in (4.1) is valid in all cases.

5. Behavior of the ARE solution near unstabilizability. It is well known that the ARE

$$A^T X + XA - XBB^T X + G = 0, \quad X \in \mathbb{R}^{n \times n}, \quad G = G^T \geq 0$$

has a unique symmetric nonnegative definite stabilizing (USNDS) solution X provided that (A, B) is stabilizable and (G, A) is detectable [16]. If detectability is lost, we can only guarantee the existence and uniqueness of a *strong* solution, that is, a symmetric nonnegative definite solution such that the spectrum of the closed-loop matrix $A - BB^T X$ lies in the closed left half-plane [19].

Let (A, B) denote an unstabilizable pair where neither A nor B is zero and consider a sequence $\{(A_k, B_k)\}_{k=0}^\infty$ of stabilizable pairs converging to (A, B) . Associate with (A, B) and each (A_k, B_k) , respectively, the ARE

$$(5.1) \quad A^T X + XA - XBB^T X + I = 0,$$

$$(5.2) \quad A_k^T X + XA_k - XB_k B_k^T X + I = 0,$$

and denote by X_k the USNDS solution to (5.2). In this section, the behavior of X_k is studied as k increases, that is, as (A_k, B_k) becomes unstabilizable. In particular, it is shown that (1) the norm of X_k goes to infinity with k , and (2) X_k^{-1} has a limit whose structure is directly related to the decomposition of (A, B) into stabilizable/unstabilizable subspaces.

First, a few technical results are recalled regarding the solution X_k to (5.2) and the ARE (5.1).

PROPOSITION 5.1. *The USNDS solutions X_k to (5.2) are uniformly positive definite in k . That is, letting λ_k denote the smallest eigenvalue of X_k , there is some strictly positive constant that uniformly bounds from below the set of λ_k 's for all k .*

Proof. See Appendix B.

PROPOSITION 5.2. *The ARE (5.1) has no nonnegative definite solution when (A, B) is unstabilizable.*

Proof. See Appendix B.

The main result of this section can now be presented.

THEOREM 5.3. *Consider a sequence of stabilizable pairs (A_k, B_k) converging to an unstabilizable pair (A, B) as $k \rightarrow \infty$. Let X_k denote the USNDS solution to (5.2) and define $Y_k = X_k^{-1}$. Then*

- (1) Y_k is the USNDS solution to the dual ARE

$$(5.3) \quad -A_k Y_k - Y_k A_k^T - Y_k^2 + B_k B_k^T = 0;$$

- (2) $\|X_k\| \rightarrow +\infty$ as $k \rightarrow +\infty$;

- (3) The sequence $\{Y_k\}_{k=1}^\infty$ converges to the strong solution Y^s of the ARE

$$(5.4) \quad -AY - YA^T - Y^2 + BB^T = 0;$$

- (4) Let U be an orthogonal state transformation bringing (A, B) to the form

$$(5.5) \quad U^T A U = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \quad U^T B = \begin{pmatrix} B_1 \\ 0 \end{pmatrix},$$

where (A_{11}, B_1) is stabilizable and A_{22} has unstable eigenvalues. Then the strong solution Y^s to (5.4) is conformably partitioned in the new basis as

$$(5.6) \quad U^T Y^s U = \begin{pmatrix} Y_{11} & 0 \\ 0 & 0 \end{pmatrix},$$

where Y_{11} is positive definite and is the USNDS solution to

$$(5.7) \quad -A_{11}Y_{11} - Y_{11}A_{11}^T - Y_{11}^2 + B_1B_1^T = 0.$$

That is, the null space of Y^s is the unstabilizable invariant subspace of (A, B) .

Proof. (1) The inverse of X_k is well defined from Proposition 5.1. Define $Y_k := X_k^{-1} > 0$. Pre- and postmultiplying (5.2) by X_k^{-1} yields (5.3). Now, rewrite (5.2) as

$$X_k A_k - X_k B_k B_k^T X_k = -A_k^T X_k - I = (-A_k^T - Y_k) X_k,$$

which implies that $-A_k^T - Y_k = X_k(A_k - B_k B_k^T X_k)X_k^{-1}$. Since X_k is stabilizing for (5.2), the matrix $A_k - B_k B_k^T X_k$ is stable; hence the closed-loop matrix $-A_k^T - Y_k$ associated with (5.3) is stable. Thus, Y_k is the USNDS solution to (5.3).

(2) By contradiction, assume that $\|X_k\|$ does not go to infinity with k . Then there is a finite limit point $X_\infty \geq 0$. Taking the limit in (5.2) for the corresponding sequence of indices k_m , it then follows that X_∞ is a nonnegative definite solution to (5.1), which contradicts Proposition 5.2.

(3) Note that $1/\|Y_k\| = \lambda_{\min}(X_k)$, where λ_{\min} stands for “smallest eigenvalue of.” From Proposition 5.1, it follows that the set $\{Y_k : k = 1, 2, \dots\}$ is bounded. Let Y_∞ be a limit point of this set. As the limit of a sequence of USNDS solutions to (5.3), Y_∞ must be a strong solution to (5.4). Equation (5.4), however, has a unique strong solution Y^s since $(-A^T, I)$ is stabilizable [19]. Consequently, Y^s is the only possible limit point for the bounded sequence $\{Y_k\}_{k=1}^\infty$, which must therefore converge to Y^s .

(4) By an orthogonal state transformation, bring the pair (A, B) to the form (5.5). Since (A_{11}, B_1) is stabilizable and (I, A_{11}) is detectable, the ARE

$$A_{11}^T X + X A_{11} - X B_1 B_1^T X + I = 0$$

has a USNDS solution X_{11} that is positive definite by Proposition 5.1. It is easily verified that $Y_{11} = X_{11}^{-1}$ is then the unique positive definite stabilizing solution to (5.7) (cf. the proof of part (1) above).

Let

$$Y := U^T \begin{pmatrix} Y_{11} & 0 \\ 0 & 0 \end{pmatrix} U.$$

Elementary algebra shows that Y is a nonnegative definite solution to (5.4). Moreover, the spectrum of

$$-A^T - Y = U^T \begin{pmatrix} -A_{11}^T - Y_{11} & 0 \\ -A_{12}^T & -A_{22}^T \end{pmatrix} U$$

lies in the closed left half-plane since A_{22} has unstable eigenvalues and $-A_{11}^T - Y_{11}$ is stable by construction of Y_{11} . Consequently, Y is a strong solution to (5.4) and by the uniqueness of the strong solution to (5.4) (see part (3) above), we conclude that $Y = Y^s$, which justifies (5.6).

The previous theorem indicates that the USNDS solution Y of (5.4) becomes singular as (A, B) becomes unstabilizable. This phenomenon is partially quantified by the following result.

THEOREM 5.4. *Let (A, B) be a stabilizable pair and Y denote the USNDS solution to (5.4). Then*

$$(5.8) \quad \lambda_{\min}(Y) \leq 2\nu_C(A, B).$$

Proof. See Appendix B.

Theorems 5.3 and 5.4 are the foundation of the bounds on μ derived in the remainder of the paper. Specifically, they suggest estimating ν in terms of the closeness to singularity of the solution Y of (5.4). In turn, such estimates will readily translate into bounds on μ using Proposition 2.3. Lower and upper bounds on μ are obtained in § 7 and 8 using this approach. Yet their derivation requires another instrumental tool, which is the behavior of μ , ν , and Y in (5.4) when shifting the spectrum of A . Technical results regarding this behavior are gathered in the next section, which may be skipped by readers mostly interested in the main results.

6. Effect of shifting the spectrum of A . When shifting the spectrum of A , the USNDS solution Y of (5.4) and the measures μ and ν exhibit a remarkable behavior, which can be exploited for the derivation of lower bounds on μ .

Define the functions ν_F^+ and ν_F^- ($F = \mathbf{R}, \mathbf{C}$) of a real variable ρ as

$$(6.1) \quad \nu_F^+(\rho) = \nu_F(A + \rho I, B), \quad \nu_F^-(\rho) = \nu_F(-A - \rho I, B).$$

Also introduce the USNDS solutions $Y^+(\rho)$ to the ARE

$$(6.2) \quad -(A + \rho I)Y^+(\rho) - Y^+(\rho)(A + \rho I)^T - Y^+(\rho)^2 + BB^T = 0,$$

and $Y^-(\rho)$ to

$$(6.3) \quad (A + \rho I)Y^-(\rho) + Y^-(\rho)(A + \rho I)^T - Y^-(\rho)^2 + BB^T = 0.$$

The reason for introducing both Y^+ and Y^- lies in identity (2.16). Specifically, $\mu(A, B)$ depends on both $\nu(A, B)$ and $\nu(-A, B)$, which are related to the smallest eigenvalues of Y^+ and Y^- , respectively, by Theorem 5.3.

The fundamental properties of μ_F , ν_F^+ , ν_F^- , Y^+ , and Y^- are summarized in the next theorem.

THEOREM 6.1. *With the definitions (6.1)–(6.3), and $F = \mathbf{R}, \mathbf{C}$,*

- (1) *For any real ρ , $\mu_F(A, B) = \mu_F(A + \rho I, B)$;*
- (2) *The functions ν_F^+ and ν_F^- are monotonically decreasing and increasing, respectively;*
- (3) *The USNDS solutions $Y^+(\rho)$ and $Y^-(\rho)$ are monotonically decreasing and increasing with ρ , respectively. Moreover,*

$$\lim_{\rho \rightarrow \pm\infty} \|Y^\mp(\rho)\| = +\infty, \quad \lim_{\rho \rightarrow \pm\infty} Y^\pm(\rho) = 0;$$

- (4) *In the case where $F = \mathbf{C}$, let $\rho^* = -\operatorname{Re} \lambda^*$, with λ^* as in (2.8). In the case where $F = \mathbf{R}$, let $\rho^* = \lambda^*$ (λ^* as in (2.10)) if $(\delta A, \delta B)$ is one-dimensional, or let ρ^* be the negative of the real part of any eigenvalue of A_{22} in (2.12) if $(\delta A, \delta B)$ is two-dimensional. Then*

$$\mu_F(A, B) = \nu_F^-(\rho^*) = \nu_F^+(\rho^*),$$

$$\text{for } \rho \leq \rho^*, \quad \mu_F(A, B) = \nu_F^-(\rho) \leq \nu_F^+(\rho),$$

$$\text{for } \rho \geq \rho^*, \quad \mu_F(A, B) = \nu_F^+(\rho) \leq \nu_F^-(\rho).$$

Proof. See Appendix C.

Another instrumental property is brought out in Theorem 6.2 below and concerns the connection between $\mu(A, B)$ and the closed-loop matrices

$$(6.4) \quad K^+(\rho) = -(A + \rho I)^T - Y^+(\rho)$$

associated with the ARE (6.2). Recall the following definition of the (complex) stability radius of a stable matrix A :

$$(6.5) \quad r_C(A) = \min \{ \|\delta A\|_F : A \in \mathbf{C}^{n \times n} \text{ and } A + \delta A \text{ unstable} \},$$

as well as the following characterization [21]:

$$(6.6) \quad r_C(A) = \min_{\alpha \in \mathbf{R}} \sigma_{\min}(A + i\alpha I).$$

THEOREM 6.2. *With the definitions (6.2), (6.4), and (6.5),*

$$(6.7) \quad \mu_C(A, B) = \min_{\substack{|\rho| \leq (1/2)\|A + A^T\|_2 \\ \rho \in \mathbf{R}}} r_C(K^+(\rho)),$$

where $\|\cdot\|_2$ denotes the spectral norm.

Proof. Rewrite (6.2) as

$$(6.8) \quad K^+(\rho)^T K^+(\rho) = (A + \rho I)(A + \rho I)^T + BB^T.$$

Now,

$$\begin{aligned} (K^+(\rho) + i\alpha I)^H (K^+(\rho) + i\alpha I) &= K^+(\rho)^T K^+(\rho) + \alpha^2 I + i\alpha(K^+(\rho) - K^+(\rho)^T) \\ &= K^+(\rho)^T K^+(\rho) + \alpha^2 I + i\alpha(A^T - A) \\ &= (A + \rho I)(A + \rho I)^T + \alpha^2 I + i\alpha(A^T - A) + BB^T \\ &= (A + (\rho + i\alpha)I)(A + (\rho + i\alpha)I)^H + BB^T \\ &= [A + (\rho + i\alpha)I, B][A + (\rho + i\alpha)I, B]^H. \end{aligned}$$

Consequently,

$$(6.9) \quad \sigma_{\min}[A + (\rho + i\alpha)I, B] = \sigma_{\min}(K^+(\rho) + i\alpha I),$$

which by taking the infimum over α provides

$$(6.10) \quad \min_{\alpha \in \mathbf{R}} \sigma_{\min}[A + (\rho + i\alpha)I, B] = \min_{\alpha \in \mathbf{R}} \sigma_{\min}(K^+(\rho) + i\alpha I) = r_C(K^+(\rho)),$$

using (6.6) for $A = K^+(\rho)$. Now, from the characterization (2.5) of μ_C , taking the infimum over ρ in (6.10) yields

$$(6.11) \quad \mu_C = \min_{\rho \in \mathbf{R}} r_C(K^+(\rho)).$$

Finally, recall from § 2 that a minimizing λ^* in (2.5) is characterized by $\lambda^* = q^{*H} A q^*$, where q^* is a complex unit vector. Therefore, from (6.9), a minimizing ρ in (6.11) is obtained as $\rho^* = -\operatorname{Re} \lambda^*$; that is, $\rho^* = -\frac{1}{2} q^{*H} (A + A^T) q^*$, whence $|\rho^*| \leq \frac{1}{2} \|A + A^T\|_2$. Thus, the range of ρ in (6.11) can be restricted to $|\rho| \leq \frac{1}{2} \|A + A^T\|_2$.

Note that Theorem 6.2 applies unchanged to

$$(6.12) \quad K^-(\rho) = (A + \rho I)^T - Y^-(\rho).$$

In fact, for any real ρ ,

$$(6.13) \quad r_C(K^+(\rho)) = r_C(K^-(\rho)),$$

since elementary algebra together with the counterpart of (6.8) for $K^-(\rho)$ shows that

$$\begin{aligned}(K^+(\rho) + i\alpha I)^H (K^+(\rho) + i\alpha I) &= [A + (\rho + i\alpha)I, B][A + (\rho + i\alpha)I, B]^H \\ &= (K^-(\rho) - i\alpha I)^H (K^-(\rho) - i\alpha I).\end{aligned}$$

The interconnection and variations of ν_C^+ , ν_C^- , μ_C , and $r_C(K^+)$ are illustrated with the following pair:

$$A = \begin{pmatrix} -0.1 & -0.3 & 0.3 & -0.2 & -8 \times 10^{-6} & 0.3 & -0.3 \\ 0.3 & -0.3 & 9 \times 10^{-2} & -1 \times 10^{-2} & 1 \times 10^{-2} & -0.2 & -0.1 \\ 6 \times 10^{-4} & 2 \times 10^{-2} & -0.2 & 0.1 & -0.1 & 0.1 & -0.2 \\ 4 \times 10^{-4} & -2 \times 10^{-3} & -2 \times 10^{-3} & -0.2 & 0.2 & -3 \times 10^{-2} & -0.2 \\ -7 \times 10^{-5} & -1 \times 10^{-4} & 1 \times 10^{-4} & 1 \times 10^{-4} & -5 \times 10^{-2} & 0.1 & -3 \times 10^{-2} \\ -8 \times 10^{-6} & -2 \times 10^{-5} & 2 \times 10^{-5} & -2 \times 10^{-5} & 1 \times 10^{-6} & -0.2 & 0.2 \\ 3 \times 10^{-6} & -1 \times 10^{-6} & -2 \times 10^{-6} & 3 \times 10^{-6} & -3 \times 10^{-6} & -2 \times 10^{-6} & 0.3 \end{pmatrix},$$

$$B = \begin{pmatrix} 0.8 \\ -0.6 \\ 9 \times 10^{-2} \\ -2 \times 10^{-2} \\ 6 \times 10^{-4} \\ -6 \times 10^{-5} \\ 3 \times 10^{-6} \end{pmatrix}.$$

Note that $\frac{1}{2}\|A + A^T\|_2 \approx 0.42$ for this particular A . The decimal logarithms of ν_C^+ , ν_C^- , μ_C , $r_C(K^+)$, and the smallest eigenvalues of $Y^+(\rho)$ and $Y^-(\rho)$ are plotted versus ρ in Fig. 1.

7. Lower bounds for μ_C . In this section, a computable lower bound for μ_C is derived in two steps. First, abstract lower bounds are obtained in Lemma 7.1 and Proposition 7.2, which involve the minimizing λ^* in (2.5). Next λ^* , which is generally

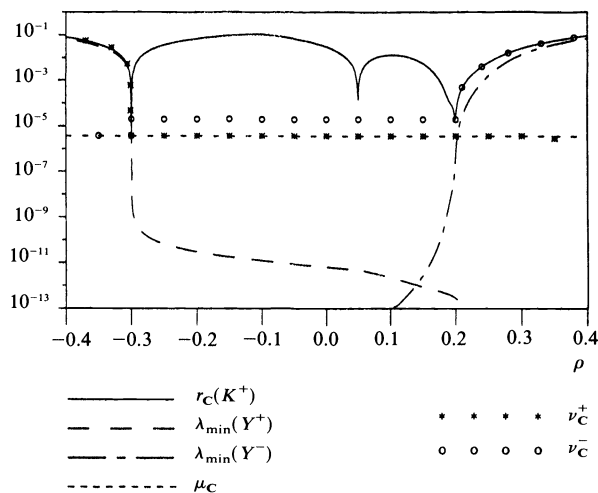


FIG. 1

unknown, is eliminated from the lower bound expression by invoking Theorem 6.2. This results in the computable lower bound (7.5) obtained in Theorem 7.4.

We begin with the derivation of abstract lower bounds.

LEMMA 7.1. *Let (λ^*, q^*) be a minimizing pair for the dual problems (2.5)–(2.6) and let $\rho^* = -\operatorname{Re} \lambda^*$. Then for $\delta\rho = \rho - \rho^* > 0$,*

$$(7.1) \quad 0 < q^{*H} Y^+(\rho) q^* \leq \frac{\mu_C^2}{\delta\rho + \sqrt{\delta\rho^2 + \mu_C^2}} \leq \frac{\mu_C^2}{2\delta\rho}.$$

Proof. Consider a unitary change of basis U , which brings the pair (A, B) to the form (2.8), and partition $U^H Y^+(\rho) U$ conformably to (2.8) as

$$\begin{pmatrix} Q & s \\ s^T & \tau \end{pmatrix}.$$

Writing the (n, n) entry of (6.2) relative to this basis, we obtain after elementary algebra

$$r^2 + \|s + \delta a\|^2 + 2 \operatorname{Re}(\lambda^* + \rho)r = \|\delta a\|^2 + \|\delta b\|^2 = \mu_C^2.$$

It follows that $r^2 + 2(\rho - \rho^*)r - \mu_C^2 \leq 0$, and thus $r \leq -\delta\rho + \sqrt{\delta\rho^2 + \mu_C^2}$. This last inequality yields (7.1) upon noting that $r = q^{*H} Y^+(\rho) q^*$ (q^* is the last basis vector in the new basis) and that $-\delta\rho + \sqrt{\delta\rho^2 + \mu_C^2} = \mu_C^2 / (\delta\rho + \sqrt{\delta\rho^2 + \mu_C^2}) \leq \mu_C^2 / (2\delta\rho)$.

Note that Lemma 7.1 applies to $Y^-(\rho)$ as defined in (6.3) by simply replacing $\rho > \rho^*$ with $\rho < \rho^*$. This suggests the following lower bound for μ_C .

PROPOSITION 7.2. *Let λ^* be a minimizing complex number in (2.5) and $\rho^* = -\operatorname{Re} \lambda^*$. Then for any $\rho \in \mathbf{R}$,*

$$(7.2) \quad \mu_C^2(A, B) \geq 2|\rho - \rho^*| \min\{\lambda_{\min}(Y^+(\rho)), \lambda_{\min}(Y^-(\rho))\}.$$

Proof. First consider the case where $\rho \geq \rho^*$. By remarking that $\lambda_{\min}(Y^+(\rho)) \leq q^{*H} Y^+(\rho) q^*$, (7.1) shows that

$$(7.3) \quad \mu_C^2(A, B) \geq 2(\rho - \rho^*) \lambda_{\min}(Y^+(\rho)).$$

Now, the counterpart of Lemma 7.1 for $Y^-(\rho)$ and $\rho < \rho^*$ yields

$$\mu_C^2(A, B) \geq 2(\rho^* - \rho) \lambda_{\min}(Y^-(\rho)),$$

which, combined with (7.3), produces (7.2).

The main obstacle to using (7.2) in practice lies in the facts that $|\rho - \rho^*|$ is unknown and that the lower bound (7.2) vanishes when $|\rho - \rho^*|$ approaches zero. This difficulty can be circumvented, however, by using an alternative lower bound when $\rho \approx \rho^*$. Such a bound is obtained from the connection between μ_C and the stability radius $r_C(K^+(\rho))$, as brought out in Theorem 6.2. The following result is needed beforehand.

LEMMA 7.3. *We have, for any real numbers ρ_1 and ρ_2 ,*

$$(7.4) \quad |r_C(K^+(\rho_1)) - r_C(K^+(\rho_2))| \leq |\rho_1 - \rho_2|.$$

Proof. This is an immediate consequence of the inequality

$$|\sigma_{\min}(A + (\rho_1 + i\alpha)I, B) - \sigma_{\min}[A + (\rho_2 + i\alpha)I, B]| \leq |\rho_1 - \rho_2|,$$

which itself follows from [11, Cor. 8.3.2, p. 286].

We can now proceed with the main theorem, which provides a computable lower bound for μ_C .

THEOREM 7.4. *For any real ρ ,*

$$(7.5) \quad \mu_C^2(A, B) \geq \gamma(\rho) r_C(K^+(\rho)),$$

where

$$\gamma(\rho) = \min \{ \frac{1}{4} r_C(K^+(\rho)), \lambda_{\min}(Y^+(\rho)), \lambda_{\min}(Y^-(\rho)) \}.$$

Proof. Recalling that $\mu_C(A, B) = r_C(K^+(\rho^*))$, where ρ^* is defined as in Proposition 7.2, Lemma 7.3 will provide an alternative lower bound on μ_C when $|\rho^* - \rho|$ vanishes. The details are as follows.

First, consider the case where $|\rho^* - \rho| \leq \frac{1}{2} r_C(K^+(\rho))$. Apply Lemma 7.3 to obtain

$$|r_C(K^+(\rho^*)) - r_C(K^+(\rho))| \leq |\rho^* - \rho| \leq \frac{1}{2} r_C(K^+(\rho)),$$

which leads to

$$\mu_C(A, B) = r_C(K^+(\rho^*)) \geq \frac{1}{2} r_C(K^+(\rho))$$

and

$$\mu_C^2(A, B) \geq \frac{1}{4} r_C^2(K^+(\rho)) \geq \gamma(\rho) r_C(K^+(\rho)).$$

Now, if, on the contrary, $|\rho^* - \rho| > \frac{1}{2} r_C(K^+(\rho))$, combine this inequality with the bound (7.2) to obtain

$$\mu_C^2(A, B) \geq 2 \{ \frac{1}{2} r_C(K^+(\rho)) \} \min \{ \lambda_{\min}(Y^+(\rho)), \lambda_{\min}(Y^-(\rho)) \} \geq r_C(K^+(\rho)) \gamma(\rho).$$

The lower bound (7.5), therefore, holds in all cases.

In practice, the bound (7.5) can easily be evaluated, the choice of ρ being the only delicate step. Since (7.5) holds for all ρ , optimal performance for the lower bound will be obtained for ρ maximizing the expression $\gamma(\rho) r_C(K^+(\rho))$. From (6.10), $r_C(K^+(\rho))$ will increase when moving away from the vertical lines of the complex plane that pass through a local minimizer of $f(\lambda) = \sigma_{\min}(A - \lambda I, B)$. Concurrently, the monotonicity properties of Y^+ and Y^- established in Theorem 6.1 indicate that the term $\lambda(\rho) := \min \{ \lambda_{\min}(Y^+(\rho)), \lambda_{\min}(Y^-(\rho)) \}$ will be maximized for $\rho = \rho_0$, where ρ_0 is uniquely characterized by

$$(7.6) \quad \lambda_{\min}(Y^+(\rho_0)) = \lambda_{\min}(Y^-(\rho_0)) = \lambda(\rho_0) := \lambda_0.$$

Now, note that for all real ρ ,

$$(7.7) \quad \lambda(\rho) = \min \{ \lambda_{\min}(Y^+(\rho)), \lambda_{\min}(Y^-(\rho)) \} \leq 2 r_C(K^+(\rho)).$$

The justification of (7.7) relies on Theorem 5.4, which provides $\lambda(\rho) \leq \lambda_{\min}(Y^+(\rho)) \leq 2 \nu_C^+(\rho)$ and on the following inequality:

$$\nu_C^+(\rho) = \min_{\operatorname{Re} \lambda \geq 0} \sigma_{\min}(A + \rho I - \lambda I, B) \leq \min_{\operatorname{Re} \lambda = 0} \sigma_{\min}(A + \rho I - \lambda I, B) = r_C(K^+(\rho)),$$

where the last identity is taken from (6.10).

Consequently, $\gamma(\rho) \approx \lambda(\rho)$ in all cases, which suggests choosing ρ close to ρ_0 defined by (7.6). However, if ρ_0 is close to the real part of some local minimizer λ_m for which $\sigma(\lambda_m) := \sigma_{\min}(A - \lambda_m I, B) \ll 1$, then $r_C(K(\rho_0)) \ll 1$ and numerical problems can be expected when computing $Y^+(\rho)$ or $Y^-(\rho)$. Specifically, if $H(\rho)$ denotes the solution to the Lyapunov equation

$$(7.8) \quad K^+(\rho)^T H(\rho) + H(\rho) K^+(\rho) + I = 0,$$

the condition of (6.2) or (6.3) degrades as $\|H(\rho)\|$ increases [15], or equivalently as $r_C(K^+(\rho))$ decreases (see [12, Thm. 2.4]). When $\|H(\rho_0)\|$ is large, it thus appears necessary to back off slightly from ρ_0 to reduce the computation sensitivity. This may seem paradoxical since, when ρ_0 is close to the real part of a local minimizer λ_m of

(2.5), $r_C(K(\rho_0))$ directly provides an estimate of the corresponding local minimum $\sigma(\lambda_m)$. However, such a situation is not as favorable as it seems because this local minimum $\sigma(\lambda_m)$ may grossly overestimate the global minimum μ_C . Moreover, evaluating $\sigma(\lambda_m)$ would involve an extra line search on the vertical line $\text{Re } \lambda = \rho_0$ since $r_C(K(\rho_0))$ cannot be accurately computed. Finally, moving away from ρ_0 increases $r_C(K^+(\rho))$ while decreasing $\lambda(\rho)$ and hence does not necessarily deteriorate the tightness of (7.5).

Summing up, the lower bound (7.5) can be optimized and reliably computed as follows. First, compute ρ_0 defined by (7.6) as well as $H(\rho_0)$ solving (7.8) for $\rho = \rho_0$. If the norm of $H(\rho_0)$ is relatively small (order of magnitude less than, say $TOL = 10^3$), proceed with (7.5). Otherwise, shift away from ρ_0 until the norm of $H(\rho)$ is within the set tolerance. The variation of $H(\rho)$ can help monitor the shift direction and magnitude, since its norm must decrease when actually moving away from a local minimizer. In most cases, one shifting step is sufficient. Note finally that the search domain for ρ_0 can be restricted to the interval $|\rho| \leq \frac{1}{2}\|A + A^T\|_2$, as shown next.

THEOREM 7.5. *Let $\rho_{\max} = \frac{1}{2}\|A + A^T\|_2$ and ρ_0, λ_0 be defined by (7.5). Then*

$$(7.9) \quad |\rho_0| \leq \rho_{\max} + \frac{\lambda_0}{2}.$$

Proof. Let u be a unit vector such that $Y^+(\rho_0)u = \lambda_0 u$. Let $\rho = \rho_0$ in (6.2). Premultiply (6.2) by u^T and postmultiply by u to get

$$-\lambda_0(2\rho_0 + u^T(A + A^T)u) - \lambda_0^2 + \|B^T u\|_2 = 0.$$

This implies that

$$\lambda_0 + 2\rho_0 + u^T(A + A^T)u \geq 0$$

and

$$\rho_0 \geq -\frac{1}{2}u^T(A + A^T)u - \frac{\lambda_0}{2} \geq -\rho_{\max} - \frac{\lambda_0}{2}.$$

The same manipulation with $Y^-(\rho_0)$ and (6.3) yields the counterpart $\rho_0 \leq \rho_{\max} + \lambda_0/2$, whence (7.9).

Gathering all the previous results and comments leads to the following algorithm to produce a lower bound on μ_C .

Algorithm 7.6

1. Find an estimate $\hat{\rho}_0$ of ρ_0 in the interval $[-\rho_{\max}, \rho_{\max}]$ (ρ_0 and ρ_{\max} as in Theorem 7.5). A Golden Section search, or an interpolation on the logarithm of the smallest eigenvalue of $Y^+(\rho)$ ($Y^-(\rho)$) can be used for this purpose.
2. Initialize with $\rho = \hat{\rho}_0$.
3. Compute $\|H(\rho)\|$.
4. If $\|H(\rho)\| > TOL$, shift ρ away from ρ_0 by $\pm 10/TOL$. Go to Step 3.
5. Else, compute $r_C(K^+(\rho))$ by a bisection method [3] and return the lower bound (7.5).

The most costly part of the algorithm is the estimation of ρ_0 , since each iteration of the search involves solving two AREs and finding the smallest eigenvalues of their solution. However, fewer than five steps are generally needed to obtain a satisfactory estimate $\hat{\rho}_0$ (see Table 2).

This section concludes with a few words on the condition of all the computations involved. The AREs are well conditioned away from the local minima of $r_C(K^+(\rho))$

and so is the estimation of $r_C(K^+(\rho))$ itself since for any stable matrix A , $|\delta r_C(A)|/r_C(A) \leq \|\delta A\|/r_C(A)$. The smallest eigenvalue of $Y^+(\rho)$ or $Y^-(\rho)$ can be computed to reasonably high relative precision, as suggested by the work in [7]. Finally, if $Y^+(\rho) + \delta Y$ denotes the computed solution to (6.2) and provided that μ_C is not too small in comparison with $\|H(\rho)\|$, the bound (7.5) remains valid with $\lambda_{\min}(Y^+(\rho) + \delta Y)$ replacing $\lambda_{\min}(Y^+(\rho))$. This assertion is justified as follows.

Suppose that a numerically stable ARE solver applied to (6.2) provides the solution $Y^+(\rho) + \delta Y$ of a nearby ARE with parameters $(A + \delta A, B + \delta B)$. Assume also that $\|(\delta A, \delta B)\|_F \ll \mu_C$, which will be the case when μ_C is not too small compared to $\|(A, B)\|_F$. Finally, denote by $\lambda_{\min}(Y^+(\rho)) + \delta \lambda$ the smallest eigenvalue of $Y^+(\rho) + \delta Y$. Consider now, below, the expression $(A + \delta A, B + \delta B)$ in the same basis with respect to which (A, B) has representation (2.8):

$$A + \delta A \equiv \begin{pmatrix} A_{11} + \delta A_{11} & a_{12} + \delta a_{12} \\ (\delta a + \delta a_{21})^H & \lambda^* + \delta \lambda^* \end{pmatrix}, \quad B + \delta B \equiv \begin{pmatrix} B_1 + \delta B_1 \\ (\delta b + \delta b_2)^H \end{pmatrix},$$

with $\|(\delta a, \delta b)\| = \mu_C$, $\|(\delta a_{21}, \delta b_2)\| \leq \|(\delta A, \delta B)\|_F$, and $|\delta \lambda^*| \leq \|\delta A\|_F$. By analogy to Lemma 7.1, writing the (n, n) entry of the perturbed ARE satisfied by $Y^+(\rho) + \delta Y$ leads to

$$(7.10) \quad \lambda_{\min}(Y^+(\rho)) + \delta \lambda \leq \frac{\|(\delta a, \delta b) + (\delta a_{21}, \delta b_2)\|^2}{2|\rho - (\rho^* + \delta \rho^*)|},$$

where $|\delta \rho^*| = |\operatorname{Re}(\delta \lambda^*)| \leq \|\delta A\|_F$. From (7.10), we obtain

$$\lambda_{\min}(Y^+(\rho)) + \delta \lambda \leq \frac{(\mu_C(A, B) + \|(\delta A, \delta B)\|_F)^2}{2(|\rho - \rho^*| - \|\delta A\|_F)} \approx \frac{\mu_C^2(A, B)}{2|\rho - \rho^*|},$$

provided that $\|(\delta A, \delta B)\|_F \ll \mu_C$ and $\|\delta A\|_F \ll \|\rho - \rho^*\|$. Therefore, bound (7.5) still applies for the computed eigenvalue $\lambda_{\min}(Y^+(\rho)) + \delta \lambda$, and the discrepancy $\delta \lambda$ can only affect its sharpness.

8. Upper bounds for μ_R . In this section, a numerical method to obtain a realistic upper bound on μ_R is described and justified from a theoretical standpoint. This method exploits the representation of (A, B) under the orthogonal coordinate transformation that diagonalizes the USNDS solution Y^+ to (5.4). The resulting upper bound and the lower bound described in § 7 are generally close to each other (within a factor of 10 to 100), and their combination thus provides a reasonable estimate of μ in most cases.

THEOREM 8.1. *Assume that (A, B) is controllable and, without loss of generality, that $\|A\|_F = 1$. Let Y^+ denote the USNDS solution to (5.4) and consider an orthogonal matrix $U = (u_1, \dots, u_n)$ such that $U^T Y^+ U = \operatorname{Diag}(\lambda_1, \dots, \lambda_n)$ with $\lambda_1 \geq \dots \geq \lambda_n$ ($\lambda_n > 0$ from Proposition 5.1). Let $(\tilde{A}, \tilde{B}) = (U^T A U, U^T B)$ denote the transformed pair and α_{ij}, b_i denote the generic entry of \tilde{A} and the i th row of \tilde{B} , respectively. Then*

$$(8.1) \quad |\alpha_{ij}| \leq \frac{\lambda_i}{\lambda_j} + \sqrt{\frac{\lambda_i}{\lambda_j} (2\alpha_{ii} + \lambda_i)(2\alpha_{jj} + \lambda_j)} \quad \text{for } i > j,$$

$$(8.2) \quad \|b_i\| = \sqrt{\lambda_i^2 + 2\lambda_i \alpha_{ii}}.$$

Proof. Write (5.4) with respect to the basis change U . Looking at the (i, i) entry, we obtain

$$-2\alpha_{ii}\lambda_i - \lambda_i^2 + b_i^T b_i = 0.$$

Identity (8.2) follows immediately. Now, consider the (i, j) entry of (5.4) for $i > j$,

$$-\lambda_j \alpha_{ij} - \lambda_i \alpha_{ji} + b_i^T b_j = 0.$$

Then

$$\alpha_{ij} = -\frac{\lambda_i}{\lambda_j} \alpha_{ji} + \frac{b_i^T b_j}{\lambda_j},$$

and (8.1) follows using (8.2) and $|\alpha_{ji}| \leq 1$ since $\|A\|_F = 1$.

If (A, B) is nearly uncontrollable, Y^+ is nearly singular ($\lambda_n \ll 1$) and with respect to the basis defined by U in Theorem 8.1, the transformed pair (\tilde{A}, \tilde{B}) has the following characteristics: (1) the last row of \tilde{B} is of very small magnitude, and (2) the entries of \tilde{A} near the bottom left corner have small magnitude as well. Such a structure is reminiscent of the form (2.8). We can thus hope that U is close to the transformation that brings (A, B) to the form (2.8). This, in turn, suggests performing the coordinate transformation U prior to running a descent algorithm. The odds of hitting irrelevant local minima and obtaining excessively optimistic estimates of μ_R are thereby significantly reduced. This idea is summarized in the following algorithm that computes an upper bound for μ_R .

Algorithm 8.2

1. Compute the USNDS solution Y^+ to (5.3), using a Schur solver.
2. Compute the eigenvalues of Y^+ and the orthogonal transformation U , which diagonalizes Y^+ with eigenvalues in decreasing order.
3. Set $\tilde{A} := U^T A U$ and $\tilde{B} := U^T B$.
4. Do a line descent (one-dimensional perturbations) initialized with (\tilde{A}, \tilde{B}) . Let μ_1 denote the local minimum encountered.
5. Do a two-dimensional descent starting with (\tilde{A}, \tilde{B}) . Let μ_2 denote the local minimum encountered.
6. Set $\mu^+ := \min(\mu_1, \mu_2)$.
7. Repeat steps 1–5 for the USNDS solution Y^- to (6.3) with $\rho = 0$. Compute the counterpart μ^- of μ^+ as in step 6.
8. Return the upper bound: $\min(\mu^+, \mu^-)$.

The benefit of this preliminary state coordinate transformation on (A, B) stems from its ability to exploit the information on directions of weakest controllability contained in the eigenstructure of Y^+ or Y^- .

9. Numerical tests. The performance of the lower and upper bounds introduced in §§ 7 and 8 has been tested on 400 nearly uncontrollable pairs (A, B) . In this experiment, the order n of the matrix ranges from 5 to 16, and various degrees of nearness to uncontrollability have been considered.

The nearly uncontrollable pairs are constructed as “perturbations” of uncontrollable pairs. Specifically,

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix},$$

where the pair obtained by zeroing the blocks A_{21} and B_2 is uncontrollable. The order k of the block A_{22} and the column dimension r of B are selected randomly between 1 and $n/2$. The entries of A_{11} , A_{12} , A_{22} , and B_1 are random numbers in $[-1/2, 1/2]$. Finally, the “perturbation” subblocks A_{21} and B_2 have the following structure:

$$(9.1) \quad A_{21} = \begin{pmatrix} \varepsilon_1 \alpha_1^T \\ \vdots \\ \varepsilon_k \alpha_k^T \end{pmatrix}, \quad B_2 = \begin{pmatrix} \varepsilon_1 b_1^T \\ \vdots \\ \varepsilon_k b_k^T \end{pmatrix},$$

where the entries of the vectors a_1, \dots, a_k and b_1, \dots, b_k are random numbers in $[-1/2, 1/2]$, and $\varepsilon_1, \dots, \varepsilon_k$ are powers of 10 ranging between 1 and 10^{-6} . The ε_i 's "control" the nearness to uncontrollability of the resulting pair and are taken either all equal or all distinct and decreasing with i . The pair (A, B) is then rescaled so that $\|A\|_F = \|B\|_F = 1$. Selecting randomly all these parameters allows the generation of pairs with very diverse levels and structures in their nearness to uncontrollability.

Note that the special structure given to the pairs (A, B) implies no loss of generality here since (i) a nearly uncontrollable pair can always be orthogonally transformed to the form (9.1) as in Propositions 2.1 and 2.2, (ii) Algorithm 7.6 works with the ARE solutions and therefore does not exploit or benefit from this initial structure, and (iii) the same remark applies for Algorithm 8.2 since it first performs a coordinate transformation that diagonalizes the ARE solution Y^\pm and replaces A and B by their expression with respect to this basis of eigenvectors. The resulting pair is thus independent of the initial choice of coordinates and consequently of the initial structure of (A, B) .

For each generated pair, we compute the ratio

$$\rho = \frac{\text{upper bound of Algorithm 8.2}}{\text{lower bound (7.5)}}.$$

This ratio must generally be greater than 1, since (7.5) bounds μ_C from below and the upper bound applies to μ_R . However, its magnitude gives an estimate of the sharpness of the bounds, since the discrepancy μ_R/μ_C does not exceed 10^{-2} in most cases, due to the choice of ε and to (4.1).

Problems of order $n = 5, 7, 10, 13, 16$ were considered, and 80 random pairs were tested for each n . The outcome of these tests is reflected in the following table which, for each n , records the percentages of samples for which ρ_I is less than 10, between 10 and 100, and over 100. See Table 1.

The average number of ARE's to be solved when computing the lower bound (step 1 of Algorithm 7.6) appears for each n in Table 2.

The numerical tests show that the bounds are relatively tight in most cases. The performance of these bounds deteriorates as n increases, mostly when the column dimension r of B is 1 (single-input case). Finally, in many cases, the upper bound of

TABLE 1

n	$\rho < 10$	$10 < \rho < 100$	$100 < \rho$
5	97	3	0
7	79	15	6
10	70	21	9
13	60	22	18
16	57	22	21

TABLE 2

n	Average # of AREs solved
5	10
7	9
10	9
13	8
16	8

Algorithm 8.2 outperforms the value of μ_R computed by standard descent methods. This suggests that the orthogonal transformation used in Algorithm 8.2 (first step) is a worthwhile preliminary step to any classical descent method.

10. Concluding remarks. Nearness to unstabilizability of a pair of matrices (A, B) implies large norm magnitude for the USNDS solution X to (5.2), or, equivalently, nearness to singularity for the solution Y to (5.4). This property is intuitive, since $\|X\|$ is associated with the cost of stabilizing system (2.1), which will be large when (A, B) is nearly unstabilizable. This connection has been exploited to derive lower bounds for the distance of (A, B) to the set of unstabilizable or uncontrollable pairs. Ways to improve the performance of the classical upper bounds on μ_R have also been suggested, which utilize the solution Y to (5.4) as well. Numerical tests have shown that these bounds provide realistic estimates of μ in most cases, especially for multi-input control. The lower bounds will be particularly useful in the robust assessment of controllability, as well as in the estimation of the condition of related problems such as pole placement (see [5]).

Note that the norm of the USNDS solution to (5.2) is somewhat related to the condition of this ARE. Specifically, the condition of (5.2) is determined by the norm of the solution H_C to the closed-loop Lyapunov equation

$$(10.1) \quad (A - BB^T X)^T H_C + H_C (A - BB^T X) + I = 0.$$

Equation (5.2) can be rewritten as

$$(10.2) \quad (A - BB^T X)^T X + X (A - BB^T X) + I + XBB^T X = 0.$$

Using the monotonicity of the solution to a stable Lyapunov equation, it follows from (10.1) and (10.2) that

$$(10.3) \quad H_C \leq X,$$

which indicates that, if the USNDS solution to (5.2) is of small norm, then it is necessarily well conditioned.

The ARE/unstabilizability connection also appears to be a powerful tool to investigate issues such as the design of state coordinate transformations that increase the distances ν and μ . Consider the following iterative coordinate transformation scheme that utilizes the USNDS solution to (5.2):

- (1) Set X_0 to be the positive definite solution to $A^T X_0 + X_0 A - X_0 B B^T X_0 + I = 0$.
- (2) For $k = 1, 2, \dots$, set $A_k = X_{k-1}^{1/2} A_{k-1} X_{k-1}^{-1/2}$, $B_k = X_{k-1}^{1/2} B_{k-1}$, and X_k to be the positive definite solution to $A_k^T X_k + X_k A_k - X_k B_k B_k^T X_k + I = 0$.

Extensive numerical tests have gathered evidence of several interesting properties for this scheme: the eigenvalues of X_k vary monotonically and converge to 1 in most cases; the norms of A_k and B_k remain of the order of those of A and B ; and the distance $\nu(A_k, B_k)$ is significantly increased after only a few steps. These empirical results suggest that such coordinate transformations would provide equivalent system representations that are more robustly stabilizable or controllable. Extension to coordinate transformations that improve the balancing of controllability and observability can also be foreseen.

The concepts developed in this paper thus give rise to a wealth of interesting open questions and offer new tools to investigate some complex minimization and optimization problems.

Appendix A.

Proof of Proposition 2.1. Consider a μ_C -minimal perturbation $(\delta A, \delta B)$. Since $(A + \delta A, B + \delta B)$ is uncontrollable, at least one eigenvalue λ of $A + \delta A$ is uncontrollable,

and there is some unit vector q such that

$$(A.1) \quad q^H A = \lambda q^H, \quad q^H B = 0.$$

Consider any matrix U whose last column vector is q . We then have

$$(A.2) \quad U^H(A + \delta A)U = \begin{pmatrix} A_{11} & a_{12} \\ 0 & \lambda \end{pmatrix}, \quad U^H(B + \delta B) = \begin{pmatrix} B_1 \\ 0 \end{pmatrix}.$$

Partition $U^H \delta A U$ and $U^H \delta B$ conformably as

$$(A.3) \quad U^H \delta A U = \begin{pmatrix} -\delta A_{11} & -\delta a_{12} \\ -\delta a^H & -\delta a_{22} \end{pmatrix}, \quad U^H \delta B = \begin{pmatrix} -\delta B_1 \\ -\delta b^H \end{pmatrix},$$

and suppose that one of the subblocks δA_{11} , δa_{12} , δa_{22} , or δB_1 in (A.3) is nonzero. Consider then $(U^H \delta \hat{A} U, U^H \delta \hat{B})$ obtained by zeroing all these subblocks. The perturbation $(\delta \hat{A}, \delta \hat{B})$ also makes (A, B) uncontrollable, since $(U^H(A + \delta \hat{A})U, U^H(B + \delta \hat{B}))$ retains the structure (A.2). From (A.3), however, $(\delta \hat{A}, \delta \hat{B})$ is of smaller Frobenius norm than the μ_C -minimal $(\delta A, \delta B)$, which provides a contradiction. Consequently, $(\delta A, \delta B)$ must have the form (2.7), and its μ_C -minimality implies that $\|\delta a\|^2 + \|\delta b\|^2 = \mu_C^2$.

Finally, we have

$$\|q^H(A - \lambda I, B)\| = \mu_C,$$

whence $\sigma_{\min}(A - \lambda I, B) \leq \mu_C$, and λ must be a minimizer in (2.5). Note that this result does not extend to ν_C , since zeroing δa_{22} in (A.3) generally does not guarantee that $(\delta \hat{A}, \delta \hat{B})$ still makes (A, B) unstabilizable.

Proof of Proposition 2.2. We adapt the argument of Proposition 2.1 to the real case. Let $(\delta A, \delta B)$ be μ_R -minimal and again consider $q \in \mathbb{C}^n$ satisfying (A.1). In the case where λ is real, q can be replaced by $\text{Re } q$ in (A.1). The matrix U can then be chosen orthogonal, and the proof of Proposition 2.1 carries through in the real field, yielding (2.9) and (2.10).

Suppose now that λ has a nonzero imaginary part. Then the subspace \mathcal{Q} spanned by the two real vectors $\text{Re } q$ and $\text{Im } q$ has dimension 2, and it is easily verified that \mathcal{Q} is $(A + \delta A)^T$ -invariant (in the real field) and that $(B + \delta B)^T \mathcal{Q} = \{\bar{0}\}$. Let (q_1, q_2) be an orthonormal basis of \mathcal{Q} and consider any orthogonal matrix U whose last two columns consist of q_1 and q_2 . Then $(U^T(A + \delta A)U, U^T(B + \delta B))$ has the structure (A.2) where λ is replaced by a real 2×2 block A_{22} . The same argument as in Proposition 2.1 then shows that $(\delta A, \delta B)$ and $(U^T A U, U^T B)$ must have the form (2.11) and (2.12), respectively. Finally, since A_{22} is real and represents the restriction of $(A + \delta A)^T$ to \mathcal{Q} , its eigenvalues are λ and $\bar{\lambda}$, which are conjugate and nonreal.

Appendix B.

Proof of Proposition 5.1. As the USNDS solution to (5.2), X_k is nonnegative definite. Let u_k be a unit eigenvector associated with λ_k . Premultiplication of (5.2) by u_k^T and postmultiplication by u_k yields

$$\lambda_k u_k^T (A_k + A_k^T) u_k - \lambda_k^2 \|B_k u_k\|^2 + 1 = 0.$$

It follows that

$$1 \leq \|B_k\|^2 \lambda_k^2 + \|A_k + A_k^T\| \lambda_k$$

and, by considering the two cases where $\lambda_k \geq 1$ and $\lambda_k < 1$, that

$$\lambda_k \geq \min \left(1, \frac{1}{\|B_k\|^2 + \|A_k + A_k^T\|} \right).$$

Consequently, the set of all λ_k 's is uniformly bounded from below by some strictly positive constant since $\{(A_k, B_k)\}_{k=0}^{\infty}$ converges to (A, B) with $B \neq 0$. The positive definiteness of each X_k is then immediate.

Proof of Proposition 5.2. If (5.1) has a solution $X \geq 0$, X must furthermore be positive definite (cf. proof of Proposition 5.1). Rewrite the ARE (5.1) as

$$(A - BB^T X)^T X + X(A - BB^T X) + XBB^T X + I = 0.$$

Since $XBB^T X + I$ is positive definite, the positive definiteness of X then implies that $A - BB^T X$ is stable by the Lyapunov theorem, which contradicts the hypothesis that (A, B) is unstabilizable.

Proof of Theorem 5.4. Denote by λ^* the complex number minimizing expression (2.4) and let q^* be a unit vector such that $\|q^{*H}(A - \lambda^* I, B)\| = \nu_C$. Note that $\operatorname{Re} \lambda^* \geq 0$. Consider any unitary change of basis such that q^* is the last basis vector in the new basis. With respect to the new basis, the pair (A, B) becomes

$$A \equiv \begin{pmatrix} A_{11} & a_{12} \\ \delta a^H & \lambda^* + \delta \alpha \end{pmatrix}, \quad B \equiv \begin{pmatrix} B_1 \\ \delta b^H \end{pmatrix},$$

with $\|\delta a\|^2 + \delta \alpha^2 + \|\delta b\|^2 = \nu_C^2$, and Y is conformably partitioned as

$$Y \equiv \begin{pmatrix} P & s \\ s^H & r \end{pmatrix}.$$

Writing the (n, n) entry of (5.4) with respect to the new coordinate system then yields

$$-(\lambda^* + \delta \alpha)r - r(\overline{\lambda^* + \delta \alpha}) - \delta a^H s - s^H \delta a - r^2 - s^H s + \delta b^H \delta b = 0$$

or, equivalently,

$$(B.1) \quad |r + \delta \alpha|^2 + (s + \delta a)^H (s + \delta a) + 2 \operatorname{Re} \lambda^* r = |\delta \alpha|^2 + \delta a^H \delta a + \delta b^H \delta b = \nu_C^2.$$

Now, r is real nonnegative and $\operatorname{Re} \lambda^* \geq 0$. Therefore, (B.1) implies that

$$|r + \delta \alpha| \leq \nu_C,$$

whence $r \leq 2\nu_C$ since $|\delta \alpha| \leq \nu_C$. The proof is complete upon noting that $\lambda_{\min}(Y) \leq r$.

Appendix C.

Proof of Theorem 6.1.

(1) Note that (A, B) uncontrollable and $(A + \rho I, B)$ uncontrollable are equivalent statements.

(2) If (A, B) is unstabilizable, then $(A + \rho I, B)$ is unstabilizable for any $\rho \geq 0$, since structure (5.5) is preserved under spectrum shifting of A , and $\rho \geq 0$ guarantees that the eigenvalues of A_{22} are shifted further into the closed right half-plane. Consequently, if a perturbation makes (A, B) unstabilizable, it also makes $(A + \rho I, B)$ unstabilizable for any $\rho \geq 0$. It follows from (6.1) and (2.2) that $\nu_F^+(A, B) \geq \nu_F^+(A + \rho I, B)$ for $\rho \geq 0$. The function ν_F^- is handled similarly.

(3) We show that $Y^+(\rho)$ is monotonically decreasing, for instance. Let $\rho_2 = \rho_1 + \delta \rho$ with $\delta \rho > 0$ and $Y^+(\rho_2) = Y^+(\rho_1) + \delta Y$. Subtracting the AREs satisfied by $Y^+(\rho_1)$ and $Y^+(\rho_2)$, we obtain after some elementary manipulation

$$-(A + \rho_1 I + Y^+(\rho_1))\delta Y - \delta Y(A + \rho_1 I + Y^+(\rho_1))^T - (2\delta \rho Y^+(\rho_2) + \delta Y^2) = 0.$$

Now $-(A + \rho_1 I + Y^+(\rho_1))$ is stable by definition of $Y^+(\rho_1)$. It follows from the Lyapunov theorem that $\delta Y < 0$ and $Y^+(\rho_2) < Y^+(\rho_1)$.

Assume that $Y^+(\rho)$ is bounded as $\rho \rightarrow -\infty$. As a monotonic function of ρ , it must then converge to some finite limit Y_∞ (see [14, p. 169]). Dividing (6.2) by ρ and taking the limit as $\rho \rightarrow -\infty$ then yields $Y_\infty = 0$. Since $Y^+(\rho)$ is decreasing and nonnegative definite, this leads to $Y^+(\rho) = 0$ for any ρ , a contradiction.

Finally, when $\rho \rightarrow +\infty$, since $Y^+(\rho)$ is monotonically decreasing and bounded from below by zero, it converges to some limit, which must be zero by the same argument as above.

(4) Consider the case where $\mathbf{F} = \mathbf{C}$, for instance. From (2.7)–(2.8), the minimal perturbation $(\delta A, \delta B)$ makes $(A + \rho^* I, B)$ unstabilizable, and its opposite makes $(-A - \rho^* I, B)$ unstabilizable. Consequently,

$$\max(\nu_c^+(\rho^*), \nu_c^-(\rho^*)) \leq \|(\delta A, \delta B)\|_F = \mu_c(A, B).$$

From Proposition 2.3, however,

$$\min(\nu_c^+(\rho^*), \nu_c^-(\rho^*)) = \mu_c(A + \rho^* I, B) = \mu_c(A, B),$$

and therefore $\nu_c^+(\rho^*) = \nu_c^-(\rho^*) = \mu_c(A, B)$. The other statements follow immediately from the monotonicity properties of ν_c^+ and ν_c^- , recalling that $\mu_c(A, B)$ bounds both functions from below.

REFERENCES

- [1] D. L. BOLEY, *Measuring how far a controllable system is from an uncontrollable one*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 249–251.
- [2] ———, *Computing rank-deficiency of rectangular matrix pencils*, System Control Lett., 9 (1987), pp. 207–214.
- [3] R. BYERS, *A bisection method for measuring the distance of a stable matrix to the unstable matrices*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 875–881.
- [4] ———, *Detecting nearly uncontrollable pairs*, MTNS Conference, Amsterdam, 1989.
- [5] J. W. DEMMEL, *On the conditioning of pole assignment*, Tech. Report No. 150, Courant Institute, New York University, New York, 1985.
- [6] ———, *A lower bound on the distance to the nearest uncontrollable system*, Tech. Report, Courant Institute, New York University, New York, 1987.
- [7] J. W. DEMMEL AND W. KAHAN, *Accurate singular values of bidiagonal matrices*, SIAM J. Sci. Stat. Comput., 11 (1990), pp. 873–912.
- [8] R. EISING, *The distance between a system and the set of uncontrollable systems*, Memo COSOR 82-19, Eindhoven University of Technology, Eindhoven, the Netherlands, 1982.
- [9] ———, *The distance between a system and the set of uncontrollable systems*, in Proc. MTNS, Beer Sheva, Israel, Springer-Verlag, New York, 1983, pp. 303–314.
- [10] ———, *Between controllable and uncontrollable*, System Control Lett., 4 (1984), pp. 263–264.
- [11] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, Johns Hopkins University Press, Baltimore, 1983.
- [12] G. HEWER AND C. KENNEY, *The sensitivity of the stable Lyapunov equation*, SIAM J. Control Optim., 26 (1988), pp. 321–344.
- [13] D. HINRICHSSEN AND A. J. PRITCHARD, *Parameterized Riccati equations and the problem of maximizing the complex stability radius*, in Proc. Workshop on The Riccati Equations in Control, Systems, and Signals, Como, Italy, 1989, pp. 136–142.
- [14] L. V. KANTOROVICH AND G. P. AKILOV, *Functional Analysis*, 2nd ed., Pergamon, New York, 1982.
- [15] C. KENNEY AND G. HEWER, *The sensitivity of the algebraic and differential Riccati equations*, SIAM J. Control Optim., 28 (1990), pp. 50–69.
- [16] V. KUCERA, *Contribution to matrix quadratic equations*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 344–347.
- [17] A. J. LAUB, *Survey of computational methods in control theory*, in Electrical Power Problems: The Mathematical Challenge, A. M. Erisman et al., eds., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1980, pp. 231–260.
- [18] C. C. PAIGE, *Properties of numerical algorithms related to computing controllability*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 130–138.

- [19] M.-A. POUBELLE, I. R. PETERSEN, M. R. GEVERS, AND R. R. BITMEAD, *Miscellany of results on an equation of Count J. F. Riccati*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 651–654.
- [20] G. W. STEWART, *Introduction to Matrix Computations*, Academic Press, New York, 1973.
- [21] C. F. VAN LOAN, *How near is a stable matrix to an unstable matrix?*, Contemp. Math., 47 (1985), pp. 465–478.
- [22] M. WICKS AND R. DECARLO, *Computing the distance to an uncontrollable system*, IEEE Trans. Automat. Control, AC-36 (1991), pp. 39–49.

WELLPOSEDNESS AND THE LAVRENTIEV PHENOMENON*

TULLIO ZOLEZZI†

Abstract. The Lavrentiev phenomenon in the calculus of variations is viewed and handled as a value Hadamard illposedness problem. Regularization is obtained by a decoupling technique of Ball-Knowles via a variational convergence approach. In this way we extend known one-dimensional results, simplifying their proofs. The same approach yields regularization for problems involving multiple integrals. A criterion for Hadamard wellposedness of multiple integrals is presented, and a new sufficient condition for non-occurrence of the one-dimensional phenomenon is obtained.

Key words. Lavrentiev phenomenon, Hadamard wellposedness, regularization, calculus of variations

AMS (MOS) subject classification. 49B50

Introduction. Consider the minimization of

$$(1) \quad I(x) = \int_a^b f(t, x, \dot{x}) \, dt$$

subject to fixed boundary conditions

$$(2) \quad x(a) = A, \quad x(b) = B.$$

Denote by $\mathcal{W}^{1,1}$ (respectively, $\mathcal{W}^{1,\infty}$) the set of all absolutely continuous (respectively, Lipschitz continuous) functions x fulfilling (2). For some choices of f and the boundary data, it happens that

$$(3) \quad \inf I(\mathcal{W}^{1,1}) < \inf I(\mathcal{W}^{1,\infty}).$$

The first example of such behavior was discovered by Lavrentiev [21]. In such cases, the global optimal value of I over various subsets H of $\mathcal{W}^{1,1}(a, b)$, whose elements fulfill (2), is a discontinuous function of H (in a sense we precisely define later). So we are faced with an ill-posed optimization problem in the sense of Hadamard (see § 1 for the definition).

Due to the Lavrentiev phenomenon (3), standard numerical methods fail to detect both the minimizers and the optimal value $\inf I(\mathcal{W}^{1,1})$ (see Ball and Knowles [4]). The interest in approximating these objects comes from optimal control (see Cesari [7]), problems in the calculus of variations with singular minimizers (see Ball and Mizel [6]), and problems in nonlinear elasticity (see Ball and Knowles [4] and Ball and Marsden [5]) involving, of course, multiple integrals with vector-valued unknown. Wellposedness theory in optimization, very briefly sketched below (see Dontchev and Zolezzi [15] for a survey) provides several useful tools for analyzing such behavior. This is the approach we emphasize in this paper. However, very few specific criteria for wellposedness in the calculus of variations are available. Further information and examples related to the Lavrentiev phenomenon can be found in [2], [4], [7], [8], [10], [11], [14], [17]–[19], [22], [27].

The aim of this paper is to employ a wellposedness analysis of the Lavrentiev phenomenon.

Two basic wellposedness notions for optimization problems are available, described below.

* Received by the editors March 12, 1990; accepted for publication (in revised form) March 1, 1991.

† Dipartimento di Matematica, Università di Genova, Via L. B. Alberti 4, 16132 Genova, Italy. The author's work was partially supported by the Ministero della Pubblica Istruzione, Italy.

Tykhonov wellposedness amounts to existence and uniqueness of the (global) optimal solution, to which every minimizing sequence converges (see [31]). Existence, uniqueness, and continuous dependence of the minimizer upon a problem's data is another wellposedness concept, which immediately reminds us of the (well-known) definition, first isolated by Hadamard, of well-posed problems in the mathematical physics (see [12, p. 227]). In many significant cases, an operator equation (or a variational inequality) is well posed in the classical sense of Hadamard if and only if an associated minimum problem has a unique optimal solution, which depends continuously on problem's data (see [24] and [26]). This basic equivalence and the analogy between the definitions involved fully justify the name *Hadamard wellposedness* for optimization problems. Such a concept is basic in the approach presented here.

In § 1 we compare Hadamard and Tykhonov wellposedness under the Lavrentiev phenomenon, showing that they are unrelated. In § 2 we consider the one-dimensional decoupling technique of Ball and Knowles [4], showing that its convergence properties are particular cases of basic results in variational convergence theory. By this approach, we obtain a new result about the regularization of the Lavrentiev phenomenon for multiple integrals with scalar unknown. Using variational convergence methods, we improve the results and simplify the proof of Ball and Knowles, avoiding ad hoc arguments and relying on a general and flexible approach to wellposedness in optimization (see Zolezzi [34]–[36]). In § 3 we obtain Hadamard wellposedness for multiple integrals, exploiting a basic link between Tykhonov and Hadamard wellposedness in the convex setting. Moreover, we present a new criterion to avoid the Lavrentiev phenomenon in the one-dimensional case.

1. Let X be a real normed space, K a fixed (nonempty) closed convex subset thereof, and $I: K \rightarrow (-\infty, +\infty]$ a proper extended real-valued function. The variational pair (K, I) will be called *value Hadamard well posed* if and only if for every sequence $H_j \subset K$ of closed sets fulfilling as $j \rightarrow +\infty$

$$(4) \quad \text{strong } \liminf H_j = K,$$

we have $\inf I(H_j) \rightarrow \inf I(K)$. Here (4) means that for every $u \in K$, we can find a sequence $u_j \in H_j$ for all j , such that $u_j \rightarrow u$ in X . Since K is weakly closed, (4) amounts to convergence of H_j to K in the sense of Mosco (see [3]). A more demanding definition of Hadamard wellposedness is introduced in § 3.

Value Hadamard wellposedness means continuous dependence of the optimal value of I upon varying constraints within K . A second fundamental concept is the following. (K, I) is *Tykhonov well posed* whenever there exists exactly one global minimizer y of I on K , and every minimizing sequence converges strongly in X to y .

Manià's example [7], [22], [27]

$$(5) \quad M(x) = \int_0^1 (x^3 - t)^2 \dot{x}^6 dt$$

exhibits value Hadamard illposedness (due to the Lavrentiev phenomenon) within

$$(6) \quad K = \{x \in \mathcal{W}^{1,1}(0, 1): x(0) = 0, x(1) = 1\},$$

since there exists some $c > 0$ such that $\inf M(K) = 0 < c \leq \inf M(H)$, where

$$H = \{x \in \mathcal{W}^{1,\infty}(0, 1): x(0) = 0, x(1) = 1\}.$$

Consider, for example, $X = \mathcal{W}^{1,1}(0, 1)$ equipped with uniform convergence and

$$H_j = \{x \in H: |\dot{x}(t)| \leq j \text{ a.e. in } (0, 1)\},$$

which fulfills (4) thanks to a standard approximation result (see [22, Thm. 1]), or H_j , the space of piecewise affine splines in K on a grid covering $[0, 1]$ with meshsize $1/j$ (see [4] and [19]). In many cases, Tykhonov and (properly defined) Hadamard wellposedness are equivalent properties (see [23] and [25] for convex optimization, and [28] for continuous functionals). The following example shows that there exist Tykhonov well-posed problems in the calculus of variations, which exhibits the Lavrentiev phenomenon.

Example. Let $y(t) = t^{1/3}$ and K be defined by (6). Consider

$$I(x) = M(x) + \int_0^1 |\dot{x} - \dot{y}| dt,$$

where M is given by (5). Of course, $I(y) = 0$; hence $\inf I(K) = 0$ while $I(x) \geq M(x) \geq c > 0$ for all $x \in H$. Let u_n be any minimizing sequence for (K, I) . Then

$$I(u_n) \geq \int_0^1 |\dot{u}_n - \dot{y}| dt$$

so that $u_n \rightarrow y$ in $\mathcal{W}^{1,1}(0, 1)$, yielding Tykhonov wellposedness of $(\mathcal{W}^{1,1}(0, 1), I)$.

2. Given an integral functional $I: K \rightarrow (-\infty, +\infty]$ of the calculus of variations that exhibits the Lavrentiev phenomenon (hence value Hadamard illposedness), we want to regularize it. Roughly speaking, we construct suitable modifications (U, J) of the original problem (K, I) (which should be properly related to it and tractable from a numerical point of view) such that, for as many sequences $L_n \subset U$ approximating K as possible, we have

$$(7) \quad \inf J(L_n) \rightarrow \inf I(K) \quad \text{as } n \rightarrow +\infty.$$

In some sense, regularization restores value Hadamard wellposedness.

We follow the decoupling approach introduced in [4], but we rely on a completely different method. The convergence property (7) will follow from the theory of variational convergence for optimization problems. This method is of general scope and can be applied to the regularization of the Lavrentiev phenomenon for multiple integrals in the calculus of variations, as shown later in this section. Finally, this method allows us to shorten and simplify some proofs of [4] and to strengthen some conclusions.

In the following, we denote subsequences as the original sequence if ambiguity does not arise. Given a sequence (K_n, I_n) of minimization problems, a sequence $u_n \in K_n$ is called *asymptotically minimizing* if and only if $\inf I_n(K_n) > -\infty$ and

$$I_n(u_n) - \inf I_n(K_n) \rightarrow 0 \quad \text{as } n \rightarrow +\infty.$$

Let $\arg \min (K, Q)$ denote the set of all global minimizers of Q over K .

The basic convergence result we exploit is shown in the following theorem.

THEOREM 1. *Let X be a real Banach space, and $I_n, Q: X \rightarrow (-\infty, +\infty]$ a given sequence. Assume that*

$$(8) \quad z_n \in X, z_n \rightharpoonup z \text{ in } X \text{ imply } \liminf I_n(z_n) \geq Q(z),$$

and that

$$(9) \quad \text{for every } \omega \in X \text{ there exists some sequence } \omega_n \in X \text{ such that } \limsup I_n(\omega_n) \leq Q(\omega).$$

Then

$$(10) \quad \limsup \inf I_n(X) \leq \inf Q(X);$$

(11) z_n asymptotically minimizing for (X, I_n) and $z_n \rightarrow z$ in X for some subsequence imply $z \in \arg \min (X, Q)$. In addition, if there exist asymptotically minimizing sequences for (X, I_n) that are weakly sequentially compact in X , then for the original sequence,

$$(12) \quad \inf I_n(X) \rightarrow \inf Q(X).$$

Theorem 1 is a particular case of [35, Thm. 1] and [36, Thms. 1 and 2].

One-dimensional case. Given f, I, A , and B as in (1), (2) put

$$(13) \quad J(x, u) = \int_a^b f(t, x, u) dt.$$

Here $x, u: [a, b] \rightarrow R^m$ for some fixed m .

Let $q \geq 1$ be fixed. We consider

$$(14) \quad \begin{aligned} L &= \{(x, \dot{x}): x \in \mathcal{W}^{1,q}(a, b), x(a) = A, x(b) = B\}, \\ K &= \{x \in \mathcal{W}^{1,q}(a, b): x(a) = A, x(b) = B\}. \end{aligned}$$

Of course, $J(x, \dot{x}) = I(x)$ whenever $x \in K$ and $I(x)$ is well defined. We are given collections of sets

$$S_h \subset K, T_h \subset L^\infty(a, b), h > 0$$

possessing the following properties. For every $y \in K$, $\omega \in L^\infty(a, b)$ and every sequence $h_n > 0$, $h_n \rightarrow 0$ there exist sequences

$$y_n \in S_{h_n}, \omega_n \in T_{h_n}$$

such that $y_n \rightarrow y$ in $\mathcal{W}^{1,q}(a, b)$, $\omega_n(t) \rightarrow \omega(t)$ almost everywhere and ω_n is uniformly bounded almost everywhere.

For given sequences h_n, ε_n put $S_n = S_{h_n}$, $T_n = T_{h_n}$, and

$$(15) \quad L_n = \left\{ (x, u) \in S_n \times T_n: \int_a^b |\dot{x} - u|^q dt \leq \varepsilon_n \right\}.$$

Using Theorem 1, we prove the following regularization result.

THEOREM 2. Assume that $f = f(t, x, u)$ is continuous and nonnegative on $[a, b] \times R^{2m}$, convex in u for all (t, x) , and satisfies

$$(16) \quad f(t, x, u) \geq \theta(|u|) \quad \text{for all } t, x, u,$$

where

$$(17) \quad \theta(s) = Cs^q + D, C > 0, \quad \text{if } q > 1;$$

θ is continuous, and $\theta(s)/s \rightarrow +\infty$ as $s \rightarrow +\infty$ if $q = 1$.

Then the following conclusions hold. For every sequence of positive numbers $\varepsilon_n \rightarrow 0$, $h_n \rightarrow 0$ we have

$$\inf J(L_n) \rightarrow \inf I(K),$$

and every asymptotically minimizing sequence for (L_n, J) possesses weak cluster points (x, \dot{x}) in $\mathcal{W}^{1,q}(a, b) \times L^q(a, b)$. Each of them satisfies $x \in \arg \min (K, I)$.

Proof. We apply Theorem 1 with

$$X = \mathcal{W}^{1,q}(a, b) \times L^q(a, b), \quad I_n = J + \text{ind } L_n, \quad Q = J + \text{ind } L,$$

where “ind” denotes the indicator function; i.e., ind L takes the value 0 on L and $+\infty$ outside L .

We begin by verifying (8). Let $z_n = (x_n, u_n) \in L_n$, $z = (u, x) \in X$ with $z_n \rightarrow z$. Since $\|\dot{x}_n - u_n\|^q \leq \varepsilon_n$ (norm in $L^q(a, b)$), we have

$$0 \leq \liminf \|\dot{x}_n - u_n\| \leq \|\dot{x} - u\|;$$

hence $\dot{x} = u$. A standard lower semicontinuity theorem ([13, Thm. 3.4, p. 74]) yields

$$\liminf I_n(z_n) = \liminf J(x_n, u_n) \geq J(x, \dot{x}) = Q(z).$$

To verify (9), let $x \in K$ and $\omega = (x, \dot{x})$. Then

$$\sup \{f(t, x(t), 0) : a \leq t \leq b\} < +\infty,$$

since x, f are continuous. By (16)

$$(18) \quad f(t, x(t), u) \geq f(t, x(t), 0), \quad a \leq t \leq b$$

whenever u is sufficiently large. For $M > 0$ consider

$$u_M(t) = \begin{cases} 0 & \text{if } |\dot{x}(t)| > M, \\ \dot{x}(t) & \text{if } |\dot{x}(t)| \leq M. \end{cases}$$

Then by (18), if M is sufficiently large,

$$(19) \quad \begin{aligned} \int_a^b f(t, x, u_M) dt &= \left(\int_{|\dot{x}| \leq M} + \int_{|\dot{x}| > M} \right) f(t, x, u_M) dt \\ &\leq \int_{|\dot{x}| \leq M} f(t, x, \dot{x}) dt + \int_{|\dot{x}| > M} f(t, x, \dot{x}) dt = J(x, \dot{x}). \end{aligned}$$

Moreover, for every n there exists M_n such that

$$(20) \quad \int_a^b |u_M - \dot{x}|^q dt = \int_{|\dot{x}| > M} |\dot{x}|^q dt < \frac{\varepsilon_n}{2},$$

provided that $M \geq M_n$. Since $v_n = u_{M_n} \in L^\infty(a, b)$, for every n there exists an almost everywhere uniformly bounded sequence $\omega_{j_n} \in T_{h_j}$ and some sequence $y_j \in S_{h_j}$ such that

$$(21) \quad y_j \rightarrow x \quad \text{in } \mathcal{W}^{1,q}(a, b), \quad \omega_{j_n} \rightarrow v_n \quad \text{a.e. as } j \rightarrow +\infty.$$

Then, by continuity of f ,

$$f(t, y_j(t), \omega_{j_n}(t)) \rightarrow f(t, x(t), v_n(t)) \quad \text{a.e.,}$$

and this convergence is dominated; hence

$$J(y_j, \omega_{j_n}) \rightarrow J(x, v_n) \quad \text{as } j \rightarrow +\infty.$$

Thus by (19), for every n ,

$$(22) \quad \limsup J(y_j, \omega_{j_n}) \leq J(x, \dot{x}).$$

By (20) and (21), given n , there exists j_n such that

$$(23) \quad \int_a^b |y_j - \omega_{j_n}|^q dt \leq \varepsilon_n \quad \text{for all } j \geq j_n.$$

Combining (22) and (23), we obtain (9).

To end the proof, by Theorem 1 we consider any asymptotically minimizing sequence (u_n, x_n) for (L_n, J) . By (10) and (16)

$$+\infty > \text{constant} \geq \int_a^b f(t, x_n, u_n) dt \geq \int_a^b \theta(|u_n|) dt,$$

so that there exists a subsequence $u_n \rightharpoonup u$ in $L^q(a, b)$. The corresponding subsequence $\dot{x}_n = \dot{x}_n - u_n + u_n \rightharpoonup u$ in $L^q(a, b)$, yielding, of course, $x_n \rightharpoonup x$ in $\mathcal{W}^{1,q}(a, b)$ with $\dot{x} = u$. Applying Theorem 1, we obtain all the required conclusions. \square

Theorem 2 extends some results of [4] in the following points. The convergence is obtained here for any asymptotically minimizing sequence, not only for minimizers of (L_n, J) . This is usually required in practice, due to approximate minimization. Moreover, there is no need to use special relations or subsequences for ε_n, h_n . More importantly, we emphasize the generality of the method used here, which gives a systematic approach to regularization based on Theorem 1 (which simplifies the proof, as compared with that of Theorem 2.1 of [4]).

In Manià's example, assumption (16) is not fulfilled. Under mild regularity conditions about the minimizing sequences, there is no loss of generality in assuming coercivity. We add a small coercifying term and then regularize as follows. Let c_r be any fixed sequence of positive numbers converging to 0 as $r \rightarrow +\infty$, and put

$$J_r(x, u) = J(x, u) + c_r \int_a^b \theta(|u|) dt.$$

THEOREM 3. *Let f satisfy the same assumptions as in Theorem 2 except (16).*

Assume that there exist a function θ and a minimizing sequence y_n for (K, I) such that

$$\int_a^b \theta(|\dot{y}_n|) dt < +\infty \quad \text{for every } n,$$

where θ is convex and fulfills (17). Then for every $\varepsilon_n, h_n \rightarrow 0$, there exists a sequence $r(n)$ increasing to $+\infty$ such that

$$(24) \quad \inf J_{r(n)}(L_n) \rightarrow \inf I(K),$$

and every weak cluster point in $\mathcal{W}^{1,q}(a, b) \times L^q(a, b)$ of any asymptotically minimizing sequence for $(L_n, J_{r(n)})$ minimizes J over L .

Proof. Since $(y_n, \dot{y}_n) \in L$, we have for every n as $r \rightarrow +\infty$

$$\begin{aligned} \limsup \inf J_r(L) &\geq \limsup J_r(y_n, \dot{y}_n) \\ &= J(y_n, \dot{y}_n) + \limsup c_r \int_a^b \theta(|\dot{y}_n|) dt = J(y_n, \dot{y}_n); \end{aligned}$$

hence $\limsup \inf J_r(L) \leq \inf J(L)$.

On the other hand, for any $(x, \dot{x}) \in L$, $J_r(x, \dot{x}) \geq J(x, \dot{x}) + c_r(\text{constant})$; hence

$$(25) \quad \inf J_r(L) \rightarrow \inf J(L) = \inf I(K).$$

Applying Theorem 2 to J_r , for every fixed r , we get $\inf J_r(L_n) \rightarrow \inf J_r(L)$ as $n \rightarrow +\infty$. Then (24) follows from [3, Cor. 1.18, p. 37]. Now, let $z_n \in L_n$ be an asymptotically minimizing sequence for $(L_n, J_{r(n)})$ such that $z_n \rightharpoonup z$ for some subsequence. Of course, $z \in L$ and, by (25), $J_{r(n)}(z_n) \rightarrow \inf I(K)$. By weak sequential lower semicontinuity

$$\liminf J_{r(n)}(z_n) \geq \liminf J(z_n) \geq J(z);$$

hence $J(z) \leq \inf I(K)$ as required. \square

Due to lack of coercivity, the existence of cluster points for asymptotically minimizing sequences in Theorem 3 cannot be guaranteed a priori. Theorem 3 can be applied to Manià's example with

$$y_0(t) = t^{1/3}, \quad \theta(s) = s^q, \quad 1 < q < \frac{3}{2}.$$

Remark. In [22] we find examples showing that the Lavrentiev phenomenon persists under coercifying perturbations, like those introduced in Theorem 3.

Abstract approach. An abstract version of the above regularization result can be obtained as follows. We are given Banach spaces Y and U , a fixed subset X of Y , a proper functional

$$J: X \times U \rightarrow (-\infty, +\infty],$$

an operator $D: X \rightarrow U$, and, for every $h > 0$, subsets $S_h \subset X$, $T_h \subset U$. We fix sequences $h_n \rightarrow 0$, $\varepsilon_n \rightarrow 0$ of positive numbers, and put

$$(x, u) \in L_n \quad \text{iff } x \in S_{h_n}, u \in T_{h_n}, \|u - Dx\| \leq \varepsilon_n;$$

$$(x, u) \in L \quad \text{iff } x \in X \quad \text{and} \quad u = Dx.$$

Then

$$\inf J(L_n) \rightarrow \inf J(L) \quad \text{as } n \rightarrow +\infty,$$

and weak cluster points of asymptotically minimizing sequences for (L_n, J) minimize J over L , provided that the following conditions hold: X is weakly sequentially closed in Y and D is weakly sequentially continuous; every sublevel set of J , i.e.,

$$\{(x, u) \in X \times U: J(x, u) \leq c\}, \quad c \in \mathbb{R},$$

is weakly sequentially closed and compact; and, finally,

for every $x \in X$ there exists a sequence of positive numbers $a_n \rightarrow 0$ such that

$$(26) \quad \{(y, u) \in L_n: J(y, u) \leq J(x, Dx) + a_n\} \neq \emptyset$$

for all n .

The proof may be obtained by mimicking that of Theorem 2. The key assumption (26) shows the role of decoupling $I(x) = J(x, Dx)$ to obtain $J(x, u)$ to enlarge the sublevel sets so as to meet every L_n , thereby fulfilling condition (9).

Multiple integrals. We are given a bounded open set Ω in \mathbb{R}^N with Lipschitz boundary, a real number $q \geq 1$, functions

$$f = f(x, u, p): \Omega \times \mathbb{R} \times \mathbb{R}^N \rightarrow \mathbb{R},$$

$u_0 \in L^\infty(\Omega) \cap \mathcal{W}^{1,q}(\Omega)$, and, for every $h > 0$, sets

$$S_h \subset u_0 + \mathcal{W}_0^{1,q}(\Omega), \quad T_h \subset L^\infty(\Omega)$$

with the following properties. For every $u \in (u_0 + \mathcal{W}_0^{1,q}(\Omega)) \cap L^\infty(\Omega)$, every $v \in L^\infty(\Omega)$, and every positive sequence $h_n \rightarrow 0$, there exist sequences $u_n \in S_{h_n}$, $v_n \in T_{h_n}$, both uniformly bounded almost everywhere in Ω , such that $u_n(x) \rightarrow u(x)$, $v_n(x) \rightarrow v(x)$ almost everywhere in Ω , and $u_n \rightarrow u$ in $\mathcal{W}^{1,q}(\Omega)$.

Put

$$I(u) = \int_{\Omega} f(x, u, \nabla u) \, dx, \quad J(u, v) = \int_{\Omega} f(x, u, v) \, dx.$$

To regularize the Lavrentiev phenomenon for $(u_0 + \mathcal{W}_0^{1,q}(\Omega), I)$ we need the following condition:

$$(27) \quad \inf \{I(u): u \in u_0 + \mathcal{W}_0^{1,q}(\Omega)\} = \inf \{I(u): u \in L^\infty(\Omega) \cap [u_0 + \mathcal{W}_0^{1,q}(\Omega)]\}.$$

A sufficient condition for (27) is given by

$$(28) \quad \begin{aligned} & \text{the functions } f(x, \pm S, 0), \quad S > 0, \text{ are equi-absolutely integrable, i.e.,} \\ & \sup \left\{ \int_E f(x, S, 0) \, dx + \int_E f(x, -S, 0) \, dx : S < 0 \right\} \rightarrow 0 \text{ as } \text{meas } E \rightarrow 0. \end{aligned}$$

To verify the above claim, let $u \in u_0 + \mathcal{W}_0^{1,q}(\Omega)$ be such that $I(u) < +\infty$, and for $S > 0$ consider $u_S \in L^\infty(\Omega)$ defined by

$$u_S(x) = u(x) \text{ if } |u(x)| \leq S, = S \text{ if } u(x) \geq S, = -S \text{ if } u(x) \leq -S.$$

Then $u_S \in u_0 + \mathcal{W}_0^{1,q}(\Omega)$ for every S sufficiently large and, as $S \rightarrow +\infty$,

$$\begin{aligned} \int_{|u| < S} f(x, u_S, \nabla u_S) \, dx &= \int_{|u| < S} f(x, u, \nabla u) \, dx \rightarrow I(u), \\ \int_{|u| \geq S} f(x, u_S, \nabla u_S) \, dx &\rightarrow 0 \end{aligned}$$

by (28). Hence $I(u_S) \rightarrow I(u)$, yielding (27).

Condition (28) is trivially fulfilled in the two-dimensional version of Manià's example [27], where

$$f(x, u, p) = (u^3 - x_1)^2 (p_1^6 + p_2^6),$$

$(x_1, x_2) \in \Omega$ if $|x_1| < 1$, $|x_2| < 1$. Sufficient conditions for (27) are also given in [16], [20, Chap. 5, Thm. 3.2], [29], [30, Thm. 6.2] and [32] (see also [1]).

For given sequences h_n, ε_n put

$$L_n = \left\{ (u, v) \in S_{h_n} \times T_{h_n} : \int_{\Omega} |\nabla u - v|^q \, dx \leq \varepsilon_n \right\}.$$

THEOREM 4. *Assume that f is nonnegative, continuous, convex in p for all (x, u) , and $f(x, u, p) \geq \theta(|p|)$ for all x, u, p , where θ fulfills (17). Moreover, suppose that (27) holds, and*

$$(29) \quad \text{for every } A > 0 \text{ there exists } C > 0 \text{ such that } f(x, u, 0) \leq C \text{ if } x \in \Omega \text{ and } |u| \leq A;$$

$$(30) \quad \text{for every } A > 0, B > 0 \text{ there exists } \phi \in L^1(\Omega) \text{ such that } f(x, u, p) \leq \phi(x) \text{ if } x \in \Omega, \\ |u| \leq A \text{ and } |p| \leq B.$$

Then for every positive sequence $\varepsilon_n \rightarrow 0$, $h_n \rightarrow 0$, we have

$$\inf J(L_n) \rightarrow \inf I[u_0 + \mathcal{W}_0^{1,q}(\Omega)],$$

and every asymptotically minimizing sequence (u_n, v_n) for (L_n, J) has cluster points $(u, \nabla u)$ with $u_n \rightharpoonup u$ in $\mathcal{W}^{1,q}(\Omega)$, $v_n \rightharpoonup \nabla u$ in $L^q(\Omega)$ for some subsequence. Each of them satisfies

$$u \in \arg \min (u_0 + \mathcal{W}_0^{1,q}(\Omega), I).$$

Proof. We apply Theorem 1 with

$$X = \mathcal{W}^{1,q}(\Omega) \times L^q(\Omega), \quad I_n = J + \text{ind } L_n, \quad Q = J + \text{ind } L,$$

where

$$L = \{(u, \nabla u) \in X : u \in u_0 + \mathcal{W}^{1,q}(\Omega)\}.$$

Condition (8) follows from a standard semicontinuity theorem [13, Thm. 3.4, p. 74]. To verify (9), fix $a > 0$ and let $\omega \in [u_0 + \mathcal{W}_0^{1,q}(\Omega)] \cap L^\infty(\Omega)$ be such that (by (27))

$$(31) \quad I(\omega) \leq \inf J(L) + a.$$

Given $M > 0$, we consider

$$v_M(x) = 0 \quad \text{if } |\nabla \omega(x)| > M, \quad v_M(x) = \nabla \omega(x) \quad \text{if } |\nabla \omega(x)| \leq M.$$

Then

$$(32) \quad J(\omega, v_M) = \int_{|\nabla \omega| > M} f(x, \omega, 0) \, dx + \int_{|\nabla \omega| \leq M} f(x, \omega, \nabla \omega) \, dx.$$

By (29) and (17), for M sufficiently large,

$$\int_{|\nabla \omega| > M} f(x, \omega, 0) \, dx \leq \int_{|\nabla \omega| > M} f(x, \omega, \nabla \omega) \, dx;$$

hence by (32)

$$(33) \quad J(\omega, v_M) = \int_{\Omega} f(x, \omega, \nabla \omega) \, dx \leq a + \inf J(L).$$

Given n , there exists $M_n > 0$ such that

$$(34) \quad \int_{\Omega} |v_{M_n} - \nabla \omega|^q \, dx = \int_{|\nabla \omega| > M_n} |\nabla \omega|^q \, dx \leq \frac{\varepsilon_n}{2},$$

and (33) holds with $M = M_n$; hence

$$J(\omega, v_{M_n}) \leq a + \inf J(L).$$

For every n there exists almost everywhere uniformly bounded sequences $u_k \in S_{hk}$, $v_k^n \in T_{hk}$ such that as $k \rightarrow +\infty$, $u_k \rightarrow \omega$ in $\mathcal{W}^{1,q}(\Omega)$ and almost everywhere in Ω , $v_k^n \rightarrow v_{M_n}$ almost everywhere in Ω .

By (34), for every n there exists k_n such that

$$(35) \quad \int_{\Omega} |\nabla u_k - v_k^n|^q \leq \varepsilon_n \quad \text{if } k \geq k_n.$$

By (30) and dominated convergence, $J(u_k, v_k^n) \rightarrow J(\omega, v_{M_n})$ for every n ; hence

$$(36) \quad \limsup J(u_k, v_k^n) \leq \inf J(L).$$

Thus (9) follows from (35) and (36). Let (u_n, v_n) be any asymptotically minimizing sequence for (L_n, J) . Then by (10) and coercivity of f , we get for some subsequence $v_n \rightharpoonup v$ in $L^q(\Omega)$; hence $\nabla u_n \rightharpoonup v$ in $L^q(\Omega)$; moreover, u_n is bounded in $L^q(\Omega)$ by Poincaré's inequality. This gives the required conclusion about cluster points of (u_n, v_n) if $q > 1$. If $q = 1$, boundedness of u_n in $\mathcal{W}^{1,1}(\Omega)$ implies that for some subsequence (via the Rellich–Kondrachov embedding theorem) $u_n \rightarrow u$ in $L^1(\Omega)$, again yielding the conclusion. \square

3. Here we find sufficient conditions for a strong form of Hadamard wellposedness of $(\mathcal{W}_0^{1,q}(\Omega), I)$, where

$$I(u) = \int_{\Omega} f(x, \nabla u) \, dx.$$

More precisely, we want to obtain that

for every sequence of nonempty closed convex subsets $K, K_j \subset \mathcal{W}_0^{1,q}(\Omega)$ such that

$$(37) \quad M\text{-}\lim K_j = K \quad \text{in } \mathcal{W}_0^{1,q}(\Omega),$$

we have, as $j \rightarrow +\infty$, $\inf I(K_j) \rightarrow \inf I(K)$; every asymptotically minimizing sequence for (K_j, I) converges strongly in $\mathcal{W}_0^{1,q}(\Omega)$ toward $\arg \min (K, I)$.

Here $M\text{-}\lim K_j = K$ denotes convergence in the sense of Mosco (see [3]), i.e., $\text{strong lim inf } K_j = K = \text{weak seq. lim sup } K_j$.

In other words, we want to detect $\inf I(K)$ and at least one minimizer of (K, I) by using only approximations K_j to K and the original integral functional I (without regularization), for all suitable K .

THEOREM 5. *Let $\Omega \subset \mathbb{R}^m$ be bounded, open, and connected. Assume that $f = f(x, p) : \Omega \times \mathbb{R}^m \rightarrow \mathbb{R}$ is a Carathéodory function with $f(x, \cdot)$ strictly convex for almost every x . Assume also that there exist functions $a(\cdot) \in L^r(\Omega)$, $b(\cdot) \in L^1(\Omega)$, where $q > 1$ and $r^{-1} + q^{-1} = 1$, and constants $C > 0$ and D such that*

$$(38) \quad C|p|^q + D \leq f(x, p) \leq a(x)|p|^q + b(x) \quad (\text{a.e. } x \in \mathbb{R}, p \in \mathbb{R}^m).$$

Then, property (37) holds. Moreover, (K, I) is Tykhonov well posed for every K , as in (37).

Proof. Fix K and let u_n be any minimizing sequence for (K, I) . By coercivity, u_n is bounded in $\mathcal{W}_0^{1,q}(\Omega)$. Hence, for some subsequence, $u_n \rightharpoonup u \in K$. By weak sequential lower semicontinuity of I , we get $u \in \arg \min (K, I)$. Strict convexity of $f(x, \cdot)$ yields uniqueness of $\arg \min (K, I)$. Thus for the original sequence we have

$$\begin{aligned} \nabla u_n &\rightharpoonup \nabla u \quad \text{in } L^q(\Omega), \\ \int_{\Omega} f(x, \nabla u_n) \, dx &\rightarrow \int_{\Omega} f(x, \nabla u) \, dx. \end{aligned}$$

By [33, Thm. 3]

$$\nabla u_n \rightarrow \nabla u \quad \text{in } L^q(\Omega);$$

hence $u_n \rightarrow u$ in $\mathcal{W}_0^{1,q}(\Omega)$, proving Tykhonov wellposedness of (K, I) . By (38), I is convex and bounded on every ball in $\mathcal{W}_0^{1,q}(\Omega)$, hence Lipschitz continuous on every bounded set (see, e.g., [9, cor., p. 35]).

Then [25, Thm. 3.1] and continuity of I yield (37). If u_j is any asymptotically minimizing sequence, then $I(u_j) \rightarrow \inf I(K)$, while for some subsequence $u_j \rightharpoonup u$ in $\mathcal{W}_0^{1,q}(\Omega)$ by (38). Hence $u \in K$ by Mosco convergence, and $u = \arg \min (K, I)$ by lower semicontinuity of I . Uniqueness of $\arg \min (K, I)$ yields $I(u_j) \rightarrow I(u)$, $u_j \rightarrow u$ for the original sequence. Again, by [33] we obtain strong convergence. \square

Remark. Tykhonov wellposedness is obtained in [5, Thm. 4.9] for multiple integrals with vector-valued unknown. Theorem 5 can be easily generalized, e.g., to

$$I(u) = \int_{\Omega} f(x, \nabla u) \, dx + \int_{\Omega} g(x, u) \, dx$$

for suitable functions g .

Under the assumptions of Theorem 5, nonoccurrence of the Lavrentiev phenomenon is, of course, trivial, due to (automatic) continuity of I .

In the next result, we return to the one-dimensional case (without convexity assumptions) by considering I defined by (1), K given by (14), and

$$H = \{x \in \mathcal{W}^{1,\infty}(a, b) : x(a) = A, x(b) = B\}.$$

We obtain the following criterion for nonoccurrence of the Lavrentiev phenomenon (hence a necessary condition for value Hadamard wellposedness).

THEOREM 6. *We have*

$$\inf I(H) = \inf I(K),$$

provided that f satisfies the following assumptions:

(39) *f is a nonnegative Carathéodory function on $[a, b] \times R^{2m}$ such that for every $C > 0$ there exists $\phi \in L^1(a, b)$ with $f(t, x, u) \leq \phi(t)$ if $|x| + |u| \leq C$;*

(40) *for every $x \in K$ with $I(x) < +\infty$, if $p_j \in R^m$ and $p_j \rightarrow 0$, if $y_j \in H$ and $y_j \rightarrow x$ in $\mathcal{W}^{1,q}(a, b)$ then, as $j \rightarrow +\infty$*

$$\int_a^b |f(t, y_j, \dot{x} + p_j) - f(t, x, \dot{x} + p_j)| dt \rightarrow 0;$$

for every $x \in K$ with $I(x) < +\infty$, there exist $C > 0$ and $\Theta \in L^1(a, b)$ such that

$$(41) \quad f(t, x(t), \dot{x}(t) + p) \leq \Theta(t)$$

for every $p \in R^m$ with $|p| \leq C$.

Proof. Given $x \in K$ such that $I(x) < +\infty$, consider

$$y_M(t) = A + \int_a^t u_M ds + (t - a)p_M, \quad M > 0,$$

where

$$p_M = \left(B - A - \int_a^b u_M dt \right) (b - a)^{-1}$$

and

$$u_M(t) = \begin{cases} 0 & \text{if } |\dot{x}(t)| > M, \\ \dot{x}(t) & \text{if } |\dot{x}(t)| \leq M. \end{cases}$$

Then $y_M \in H$ for every M . Moreover,

$$p_M \rightarrow 0, \quad y_M \rightarrow x \quad \text{in } \mathcal{W}^{1,q}(a, b) \quad \text{as } M \rightarrow +\infty.$$

By (39), $\int_{|\dot{x}| > M} f(t, y_M, p_M) dt \rightarrow 0$. Moreover,

$$\begin{aligned} & \int_{|\dot{x}| \leq M} f(t, y_M, \dot{y}_M) dt \\ & \leq \int_{|\dot{x}| \leq M} [f(t, y_M, \dot{x} + p_M) - f(t, x, \dot{x} + p_M)] dt + \int_a^b f(t, x, \dot{x} + p_M) dt. \end{aligned}$$

Hence, as $M \rightarrow +\infty$,

$$\limsup I(y_M) = \limsup \int_{|\dot{x}| \leq M} f(t, y_M, \dot{y}_M) dt \leq \int_a^b f(t, x, \dot{x}) dt$$

by (40) and (41). This yields the conclusion. \square

Remark. Assumption (40) is violated in Manià's example; it suffices to take $x = 0 = p_i$, $y_j(t) = tj^{2/3}$ if $0 \leq t \leq 1/j$, $y_j(t) = t^{1/3}$ if $t \geq 1/j$.

The next example shows that (40) and (41) are independent of the key condition required by Angell [2] in his criterion for nonoccurrence of the Lavrentiev phenomenon. This condition (in a somewhat more general form, see [7, p. 509]) requires that as $n \rightarrow +\infty$

$$\int_a^b |f(t, x_n, \dot{x}_n) - f(t, x, \dot{x}_n)| dt \rightarrow 0$$

for every sequence $x_n \in \mathcal{W}^{1,1}(a, b)$ such that $x_n \rightarrow x$ uniformly on $[a, b]$.

Example. Let $f(t, x, u) = |u|^3(1 + |x|)^{-1}$ and

$$I(x) = \int_0^1 |\dot{x}|^3(1 + |x|)^{-1} dt, \quad x(0) = 0, \quad x(1) = 0.$$

Then $|f(t, y_j, \dot{x} + p_j) - f(t, x, \dot{x} + p_j)| \leq (\text{constant})(|\dot{x}|^3 + |p_j|^3)|x - y_j|$; hence (40) holds with $q = 3$. Moreover, (41) is satisfied since $f(t, x, \dot{x} + p) \leq (\text{constant})(|\dot{x}|^3 + |p|^3)$. We show that Angell's basic condition is not satisfied. Consider

$$x_n(t) = \begin{cases} t\sqrt{n} & \text{if } 0 \leq t \leq 1/2n, \\ -t\sqrt{n} + \frac{1}{\sqrt{n}} & \text{if } 1/2n \leq t \leq 1/n, \\ 0 & \text{if } 1/n \leq t \leq 1, \end{cases}$$

and put $x(t) = 0$. Then, for all n ,

$$\int_0^1 |f(t, x_n, \dot{x}_n) - f(t, x, \dot{x}_n)| dt \geq n\sqrt{n} \int_0^{1/n} \frac{x_n dt}{1 + x_n} \geq \frac{1}{6}.$$

Remark. Another set of conditions precluding the Lavrentiev phenomenon (provided that $f(t, \cdot, u)$ is locally Lipschitz) are presented in [22, Thm. 3]. In particular, condition (b) together with (41) imply (40).

Acknowledgments. We thank L. Boccardo, who brought to our attention [5] and showed us the very simple proof leading to Tykhonov wellposedness in § 3. A preliminary version of this paper was presented at the Workshop on Wellposedness and Stability of Optimization Problems and Related Topics (Sofia, Bulgaria, March 1989). Thanks are also due to a referee and to the editors for editorial assistance.

REFERENCES

- [1] L. AMBROSIO, O. ASCENZI, AND G. BUTTAZZO, *Lipschitz regularity for minimizers of integral functionals with highly discontinuous integrands*, J. Math. Anal. Appl., 142 (1989), pp. 301–316.
- [2] T. S. ANGELL, *A note on approximation of optimal solutions of free problems of the calculus of variations*, Rend. Circ. Mat. Palermo, 28 (1979), pp. 258–272.
- [3] H. ATTOUCH, *Variational Convergence for Functional and Operators*, Pitman, Boston, 1984.
- [4] J. BALL AND G. KNOWLES, *A numerical method for detecting singular minimizers*, Numer. Math., 51 (1987), pp. 181–197.
- [5] J. M. BALL AND J. E. MARSDEN, *Quaxiconvexity at the boundary, positivity of the second variation and elastic stability*, Arch. Rational Mech. Anal., 86 (1984), pp. 251–277.
- [6] J. BALL AND V. MIZEL, *One-dimensional variational problems whose minimizers do not satisfy the Euler–Lagrange equation*, Arch. Rational Mech. Anal., 90 (1985), pp. 325–388.
- [7] L. CESARI, *Optimization Theory and Applications*, Springer-Verlag, Berlin, New York, 1983.
- [8] L. CESARI AND T. ANGELL, *On the Lavrentiev phenomenon*, Calcolo, 22 (1985), pp. 17–29.
- [9] F. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.

- [10] F. H. CLARKE AND P. D. LOEWEN, *An intermediate existence theory in the calculus of variations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 16 (1989), pp. 487–526.
- [11] F. CLARKE AND R. VINTER, *Regularity properties of solutions to the basic problem in the calculus of variations*, Trans. Amer. Math. Soc., 289 (1985), pp. 73–98.
- [12] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, Vol. II, Interscience, New York, 1962.
- [13] B. DACOROGNA, *Direct Methods in the Calculus of Variations*, Springer-Verlag, Berlin, New York, 1989.
- [14] A. DAVIE, *Singular minimizers in the calculus of variations in one dimension*, Arch. Rational Mech. Anal., 101 (1988), pp. 161–177.
- [15] A. DONTCHEV AND T. ZOLEZZI, *Well Posed Optimization Problems*, to appear.
- [16] P. HARTMAN AND G. STAMPACCHIA, *On some nonlinear elliptic differential-functional equations*, Acta Math., 115 (1966), pp. 271–310.
- [17] A. HEINRICHER AND V. MIZEL, *The Lavrentiev phenomenon for invariant variational problems*, Arch. Rational Mech. Anal., 102 (1988), pp. 57–93.
- [18] ———, *A new example of the Lavrentiev phenomenon*, SIAM J. Control Optim., 26 (1988), pp. 1490–1502.
- [19] G. KNOWLES, *Finite element approximation to singular minimizers, and application in non-linear elasticity*, Lecture Notes in Math., 1285 (1987), pp. 236–247.
- [20] O. A. LADYZHENSKAJA AND N. N. URALTSEVA, *Linear and Quasi Linear Elliptic Equations*, Academic Press, New York, 1968.
- [21] M. LAVRENTIEV, *Sur quelques problemes du calcul des variations*, Ann. Mat. Pura Appl., 4 (1927), pp. 7–28.
- [22] PH. LOEWEN, *On the Lavrentiev phenomenon*, Canad. Math. Bull., 30 (1987), pp. 102–108.
- [23] R. LUCCHETTI, *Some aspects of the connection between Hadamard and Tykhonov wellposedness of convex programs*, Boll. Un. Mat. Ital., IC (1982), pp. 337–345.
- [24] R. LUCCHETTI AND F. PATRONE, *A characterization of Tyhonov wellposedness for minimum problems with application to variational inequalities*, Numer. Funct. Anal. Optim., 3 (1981), pp. 461–476.
- [25] ———, *Hadamard and Tykhonov well posedness of a certain class of convex functions*, J. Math. Anal. Appl., 88 (1982), pp. 204–215.
- [26] ———, *Some properties of “well-posed” variational inequalities governed by linear operators*, Numer. Funct. Anal. Optim., 5 (1982–1983), pp. 349–361.
- [27] B. MANIÀ, *Sopra un esempio di Lavrentieff*, Boll. Un. Mat. Ital., 13 (1934), pp. 147–153.
- [28] J. REVALSKI, *Generic wellposedness in some classes of optimization problems*, Acta Univ. Carolin.—Math. Phys., 28 (1987), pp. 116–125.
- [29] R. SCHIANCHI, *An estimate for the minima of the functionals of the calculus of variations*, Differential Integral Equations, 2 (1989), pp. 326–332.
- [30] G. STAMPACCHIA, *On some regular multiple integral problems in the calculus of variations*, Comm. Pure Appl. Math., 16 (1963), pp. 383–421.
- [31] A. N. TYKHONOV, *On the stability of the functional optimization problem*, U.S.S.R. Comput. Math. and Math. Phys., 6 (1966), pp. 26–33.
- [32] G. TALENTI, *Boundedness of minimizers*, Hokkaido Math. J., 19 (1990), pp. 259–279.
- [33] A. VISINTIN, *Strong convergence results related to strict convexity*, Comm. Partial Differential Equations, 9 (1984), pp. 439–466.
- [34] T. ZOLEZZI, *Some approximation and convergence problems in optimization*, Serdica, 9 (1983), pp. 400–406.
- [35] ———, *On stability analysis in mathematical programming*, Math. Programming Study, 21 (1984), pp. 227–242.
- [36] ———, *Stability analysis in optimization*, Lecture Notes in Math., 1190 (1986), pp. 397–429.

PARAMETER CONVERGENCE AND UNIQUENESS IN NONMINIMAL PARAMETERIZATIONS FOR MULTIVARIABLE ADAPTIVE CONTROL*

YOSEF WILLNER[†], MICHAEL HEYMANN[‡], AND MARC BODSON[§]

Abstract. The issue of parameter convergence in multivariable adaptive control is addressed in a general framework. Parameter convergence is guaranteed if a certain design identity has a unique solution and if the inputs satisfy persistency of excitation conditions. The uniqueness of the solution of the design identity can be obtained, in general, by using parameterizations that, although nonminimal, are structured so as to guarantee uniqueness. This concept is illustrated with a direct adaptive pole placement algorithm, which is modified to guarantee uniqueness, and it is shown how the results can be used to establish stability and convergence properties of the algorithm.

Key words. adaptive control, multivariable systems, pole placement, parameter convergence, persistency of excitation, parameterizations

AMS(MOS) subject classifications. 93C40, 93C35, 93B55, 93D21, 93B10

1. Introduction. The issue of parameter convergence in adaptive control has received some attention in recent years (see, among others, [1] and [2] in discrete-time and [3] and [4] in continuous-time). It was found that several single-input single-output (SISO) schemes possessed exponential parameter convergence properties, provided that persistency of excitation (or sufficient richness) conditions were satisfied.

Although it is often argued that parameter convergence is not necessary in adaptive control (boundedness and tracking being the only objectives of model reference adaptive control, for example), there are important reasons to study this problem. First, exponential stability guarantees a certain degree of robustness (cf. [4], [5]). In the presence of noise, adaptive schemes exhibit parameter drift and a burst phenomenon (cf. [6], [17]), which can be avoided if persistency of excitation conditions are met. The problem can also be avoided using deadzones and projections, but only at the cost of additional prior information.

Another advantage of parameter convergence is that the closed-loop system actually has the asymptotic properties for which the controller was designed. Indeed, consider the case of a model reference adaptive scheme with an input signal that is constant over a long period of time. While the tracking error converges to zero, the closed-loop poles may converge to arbitrary locations. This may result in large transients when the reference input later varies.

It is important to note that we address ourselves here to a strong form of parameter convergence, namely uniform *exponential* parameter convergence to the *nominal* values of the parameters (also called correct, or true values). This form of convergence requires conditions of persistency of excitation. Weaker conditions on the input signals result in weaker forms of parameter convergence. For example, it is known that the parameters of the recursive least-squares algorithm converge without further conditions than those needed for stability [18]. In that case, however, the parameters do not necessarily converge to their nominal values, and the convergence is not exponential in general.

* Received by the editors May 14, 1990; accepted for publication (in revised form) April 17, 1991.

[†] Department of Electrical Engineering, Technion, Haifa 32000, Israel.

[‡] Department of Computer Science, Technion, Haifa 32000, Israel. This author was supported by the Technion Fund for Promotional Research.

[§] Department of Electrical and Computer Engineering, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213. This author was supported by National Science Foundation grant ECS 88-10145 and by the Lady Davis Foundation at the Technion, Israel.

The advantages of the strong form of parameter convergence mentioned above, as far as robustness and asymptotic performance are concerned, are lost in such cases, where the persistency of excitation conditions are relaxed.

Very few rigorous proofs of stability have been published for multivariable adaptive control algorithms, and parameter convergence has not been established. In fact, parameter convergence often *cannot* be guaranteed for the existing schemes, even with sufficiently rich inputs. This happens because the parameterizations are not *unique*. In a model reference adaptive control algorithm, for example, this means that an infinite number of values of the parameters exist such that model matching is achieved.

In the context of recursive identification, it was shown [7] that, using unique parameterizations, frequency-domain conditions on the inputs could be specified under which parameter convergence was guaranteed. The parameterization used there had the additional advantage of being *minimal*, i.e., of requiring the minimal number of parameters necessary to describe the class of systems under consideration. In direct adaptive control, minimality is rarely achieved (even in the SISO case), but it was shown in [1] that it is only necessary for the parameterization to be unique (rather than minimal) to guarantee parameter convergence under suitable persistency of excitation conditions.

Contributions of the paper. The first contribution of this paper is to extend the results of [1] to the multivariable case (§§ 2–4), thereby establishing a general framework for the convergence analysis of a large class of adaptive control algorithms. Specifically, the results show that parameter convergence is related to the *uniqueness* of the solution of a certain design identity. This result is important because it provides a criterion to guarantee parameter convergence and, furthermore, indicates that minimality is not itself necessary. For example, a SISO linear time-invariant system can be described by $2n$ parameters, where n is the order of the system. However, parameter convergence can be achieved with a pole placement algorithm with $4n$ parameters. This is obtained by giving sufficient *structure* to the nonminimal model, so that uniqueness is guaranteed. This paper proves that the same principle holds true for multivariable systems, and it gives a general framework in which to test the requirements for making exponential convergence of the parameters to their nominal values possible.

The second contribution of the paper is to show how uniqueness can be guaranteed in a specific adaptive pole placement scheme and to prove the stability and convergence properties of this scheme by applying the general results (§§ 5 and 6). As noted above, existing direct adaptive control schemes do not guarantee uniqueness in the design identity. However, we show how the adaptive pole placement scheme of [8] can be modified to achieve this result.

Two nontrivial modifications are incorporated in the scheme of [8]: the first consists in restricting the column degrees of the elements of some polynomial matrices, considering the knowledge of the observability indices of the plant. It is interesting to observe the similarity to the situation that arises in recursive parametric identification. There, the knowledge of the observability indices can be used to constrain the column degrees of a left matrix fraction description of the plant, thereby leading to a canonical representation and to the uniqueness of the parameterization. In fact, an interesting feature of a proof presented in this paper is to show how known results on canonical forms can be used to guarantee the uniqueness of a direct adaptive control parameterization (that is, the uniqueness of the solution of the corresponding design identity).

As opposed to the situation in identification, the constraint on the column degrees is insufficient to guarantee uniqueness for the adaptive pole placement algorithm. It

is found that the order of a certain observer polynomial matrix in [8] must also be modified to guarantee uniqueness of the solution of the design identity. This second modification is not obvious and is rather technical, but is given in this paper.

We follow, as much as possible, the notation and terminology of [1] and [8], to which this work is most closely related. The reader may wish to consult [1] in particular, for motivation and for additional details on some of the techniques used in this paper.

2. A general parameter estimation problem. We consider discrete-time, linear time-invariant systems modeled by the state equations

$$(2.1) \quad \begin{aligned} x(t+1) &= A_s x(t) + B_s u(t), \\ y(t) &= C_s x(t) + E_s u(t), \end{aligned}$$

where $u(t)$ is the $(m \times 1)$ input vector, $y(t)$ the $(p \times 1)$ output vector, and $x(t)$ the $(n \times 1)$ state vector. We assume that system (2.1) is minimal. Let the controllability indices μ_i , $1 \leq i \leq m$ and the observability indices ν_i , $1 \leq i \leq p$ be defined as usual (cf. [9]). Let $\mu = \max_{1 \leq i \leq m} (\mu_i)$ the maximal controllability index, simply called the controllability index, and $\nu = \max_{1 \leq i \leq p} (\nu_i)$, called the observability index. Such a system can also be represented by the right matrix fraction description

$$(2.2) \quad P(D) \cdot \xi(t) = u(t), \quad y(t) = R(D) \cdot \xi(t),$$

where $R(D)$ and $P(D)$ are $(p \times m)$ and $(m \times m)$ real polynomial matrices in the unit delay operator (i.e., $D^k x(t) = x(t-k)$). Matrices $R(D)$ and $P(D)$ exist that have the following properties (cf. [9]):

- (a) $\partial_{ej} R(D) \leq \mu_j$ and $\partial_{ej} P(D) = \mu_j$, where $\partial_{ej}[\cdot]$ denotes the maximal polynomial degree in the j th column,
- (b) $P(0)$ is nonsingular,
- (c) The matrices $R(D)$ and $P(D)$ are right coprime.

The matrix $P(D)$ can be further constrained, in particular, so that it is in some canonical form (cf. [10]). This will be discussed in § 5.

Structured nonminimal model. To introduce a general framework for the study of direct adaptive control algorithms, we replace the minimal model (2.2) by a *structured nonminimal model* of the form

$$(2.3) \quad \left[C(D) + \sum_{j=1}^{m_a} A_j(D) \alpha_j \right] y(t) = \left[E(D) + \sum_{j=0}^{m_b} B_j(D) \beta_j \right] u(t),$$

where m_a and m_b are positive integers, $\alpha_j \in \mathbb{R}^{r \times p}$, $\beta_j \in \mathbb{R}^{r \times m}$, $C(D)$ and $E(D)$ are $(r \times p)$ and $(r \times m)$ polynomial matrices with maximal degree l . $A_j(D)$ and $B_j(D)$ are $(r \times r)$ polynomial matrices of the form

$$(2.4) \quad A_j(D) = \text{diag} [a_{ij}(D)], \quad a_{ij}(D) = \sum_{k=1}^l a_{ijk} D^k, \quad a_{ijk} \in \mathbb{R},$$

$$(2.5) \quad B_j(D) = \text{diag} [b_{ij}(D)], \quad b_{ij}(D) = \sum_{k=0}^l b_{ijk} D^k, \quad b_{ijk} \in \mathbb{R}.$$

The structured nonminimal model defined by (2.3)–(2.5) is an extension of the SISO model of [1]. There, it was shown that the simplified model was adequate to describe several adaptive control algorithms. In § 5, we will show that the multivariable adaptive pole placement algorithm fits into the generalized framework. The integer r is equal

to m , the number of inputs, in that case. In other cases, it may take different values (for example, (2.3) may represent a left matrix fraction description, with $r = p$).

At this point, we let r and l be arbitrary integers, but it is assumed that, for $i = 1, 2, \dots, r$, the polynomials $\{a_{ij}(D)\}_{j=1}^{m_a}$ are linearly independent over the reals. It is assumed similarly that the $\{b_{ij}(D)\}_{j=0}^{m_b}$ are linearly independent. These assumptions imply that $\max\{m_a, m_b\} \leq l$.

We assume that the plant can be represented by the model (2.3) for some given matrices $C(D)$, $E(D)$, $A_j(D)$, and $B_j(D)$. We wish to use the model (2.3) to uniquely estimate the elements of the matrices α_j , $1 \leq j \leq m_a$ and β_j , $0 \leq j \leq m_b$ from the plant input-output data. Clearly, (2.3) constitutes a model for the plant (2.2) if and only if the following *design identity* is satisfied:

$$(2.6) \quad \left[C(D) + \sum_{j=1}^{m_a} A_j(D)\alpha_j \right] R(D) = \left[E(D) + \sum_{j=0}^{m_b} B_j(D)\beta_j \right] P(D).$$

Clearly, the elements of the matrices α_j and β_j in (2.3) can be uniquely estimated only if (2.6) has a unique solution $\{\alpha_1, \dots, \alpha_{m_a}, \beta_0, \dots, \beta_{m_b}\}$. Conversely, whenever (2.6) has a unique solution, we will show that the solution can be obtained by a direct estimation algorithm, which is exponentially convergent. This will be the focus of the ensuing discussion.

Remark 1. We wish to emphasize that the problem of finding conditions that ensure that (2.6) has a unique solution is quite different when the plant (2.2) is SISO and when the plant is MIMO (multi-input multi-output); see the following:—*In the SISO case*, for given polynomials $C(D)$ and $E(D)$, a solution to (2.6) exists if the polynomials $R(D)$ and $P(D)$ are coprime and if the degrees of the polynomials $A(D) \triangleq \sum A_j(D)\alpha_j$ and $B(D) \triangleq \sum B_j(D)\beta_j$ are sufficiently large. Among all solutions of (2.6), there is a unique solution $\{A(D), B(D)\}$ with minimal degree. Therefore, to ensure a unique solution of (2.6), it is sufficient to bound the degrees of the polynomials $A_j(D)$, $1 \leq j \leq m_a$ and $B_j(D)$, $0 \leq j \leq m_b$.

In the MIMO case, (2.6) has a solution if the matrices $R(D)$ and $P(D)$ are right coprime and if the degrees of the elements of the matrices $A(D)$ and $B(D)$ are sufficiently large. However, to ensure the uniqueness of the solution, it is necessary to restrict the maximal and the minimal powers in D of each element of $A(D)$ and $B(D)$. In other words, we can guarantee uniqueness by restricting the maximal degree of each element in $A_j(D)$ and $B_j(D)$ and by choosing some of the elements of α_j and β_j as zero. This fact will be made clearer in § 5.

For parameter estimation purposes, it is convenient to write every row of (2.3) as an independent equation

$$(2.7) \quad \left[C_i(D) + \sum_{j=1}^{m_a} a_{ij}(D)\alpha_j \right] y(t) = \left[E_i(D) + \sum_{j=0}^{m_b} b_{ij}(D)\beta_j \right] u(t)$$

for $i = 1, 2, \dots, r$, where $C_i(D)$, $E_i(D)$, α_j , and β_j are the i th row of the matrices $C(D)$, $E(D)$, α_j , and β_j , and where the polynomials $a_{ij}(D)$ and $b_{ij}(D)$ are defined in (2.4) and (2.5). These equations can be written as regression equations

$$(2.8) \quad \bar{\phi}_i^T(t) \bar{\theta}_i^* = E_i(D)u(t) - C_i(D)y(t), \quad i = 1, 2, \dots, r,$$

where

$$(2.9) \quad \bar{\phi}_i^T(t) = [a_{i1}(D)y^T(t), \dots, a_{im_a}(D)y^T(t), -b_{i0}(D)u^T(t), \dots, -b_{im_b}(D)u^T(t)],$$

$$(2.10) \quad \bar{\theta}_i^* = [\alpha_{i1}, \dots, \alpha_{im_a}, \beta_{i0}, \dots, \beta_{im_b}]^T.$$

As indicated in Remark 1, (2.6) has a unique solution, provided that the elements of $A(D) \triangleq \sum A_j(D)\alpha_j$ and $B(D) \triangleq \sum B_j(D)\beta_j$ satisfy some degree conditions. These conditions depend on the adaptive control problem, and we will give specific conditions in the case of the pole placement algorithm in § 5. The degree conditions imply that some elements in the vectors $\bar{\theta}_i^*$, $1 \leq i \leq r$ are zero. We delete these zero elements from $\bar{\theta}_i^*$, as well as the corresponding elements from $\bar{\phi}_i(t)$, and define the resulting vectors θ_i^* and $\phi_i(t)$, respectively. Equation (2.8) is then equivalent to

$$(2.11) \quad \phi_i^T(t)\theta_i^* = E_i(D)u(t) - C_i(D)y(t), \quad i = 1, 2, \dots, r.$$

Standard estimation procedures, such as the recursive least squares (RLS) algorithm, can be used to estimate each of the parameter vectors θ_i^* using input-output data of the plant. It is well known (cf. [2]) that to ensure the global convergence of the estimation algorithms, it is necessary to satisfy a persistency of excitation condition. In § 3, we will introduce linear systems called the *associated-signal systems* of (2.11). Through the use of these systems, we will show in § 4 how the persistency of excitation condition can be satisfied.

3. The associated-signal system and its output reachability. For each equation in (2.11), we define the associated-signal system, which is a linear system in state-space form. Its input vector is $u(t)$ (the input vector of the plant (2.1) or (2.2)), and $\phi_i(t)$ is its output vector. Let the state of the associated-signal system be defined as the following $(m(l+\mu))$ vector:

$$(3.1) \quad x_a(t) = [\xi^T(t-1), \xi^T(t-2), \dots, \xi^T(t-l-\mu)]^T,$$

where $\xi(t)$ is defined in (2.2), l is the same as in (2.4) and (2.5), and μ is the controllability index of the plant (2.1). The matrices $R(D)$ and $P(D)$ (in (2.2)) can be written as

$$(3.2) \quad R(D) = \sum_{k=0}^{\mu} R_k D^k, \quad P(D) = \sum_{k=0}^{\mu} P_k D^k.$$

It follows from the first equation of (2.2) that $x_a(t)$ satisfies the discrete-time state equation

$$(3.3) \quad x_a(t+1) = Ax_a(t) + Bu(t),$$

where

$$(3.4) \quad A = \begin{bmatrix} -P_0^{-1}P_1 - P_0^{-1}P_2 \cdots - P_0^{-1}P_{\mu} & 0 & \cdots & 0 \\ & \vdots & & \\ & I_{m(l+\mu-1)} & & \\ & & & 0 \end{bmatrix}, \quad B = \begin{bmatrix} P_0^{-1} \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

By using (2.2), (2.4), and (2.5), it can be shown that the vectors $\bar{\phi}_i(t)$ (in (2.9)) satisfy the equations

$$(3.5) \quad \bar{\phi}_i(t) = \bar{C}_i x_a(t) + \bar{E}_i u(t), \quad i = 1, 2, \dots, r,$$

where \bar{C}_i is the following $((pm_a + m(m_b + 1)) \times (m(l + \mu)))$ matrix:

$$(3.6) \quad \bar{C}_i = \begin{bmatrix} a_{i11}R_0 & a_{i12}R_0 + a_{i11}R_1 & \cdots & a_{i1l}R_\mu \\ \vdots & \vdots & & \vdots \\ a_{im_a1}R_0 & a_{im_a2}R_0 + a_{im_a1}R_1 & \cdots & a_{im_al}R_\mu \\ -b_{i01}P_0 & -b_{i02}P_0 - b_{i01}P_1 & \cdots & -b_{i0l}P_\mu \\ \vdots & \vdots & & \vdots \\ -b_{im_b1}P_0 & -b_{im_b2}P_0 - b_{im_b1}P_1 & \cdots & -b_{im_b l}P_\mu \end{bmatrix}.$$

The element of \bar{C}_i in the j th row and k th column is the coefficient of D^k in $a_{ij}(D)R(D)$ for $j = 1, \dots, m_a$. Similarly, for $j = 0, \dots, m_b$ the element in the $(j + m_a + 1)$ th row and k th column is the coefficient of D^k in $-(b_{ij}(D) - b_{ij}(0))P(D)$. \bar{E}_i is then the $((pm_a + m(m_b + 1)) \times m)$ matrix

$$(3.7) \quad \bar{E}_i = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ -b_{i00} \cdot I \\ -b_{i10} \cdot I \\ \vdots \\ -b_{im_b0} \cdot I \end{bmatrix},$$

where the elements a_{ijk} and b_{ijk} are defined in (2.4) and (2.5).

By definition, the *associated-signal system* of the i th equation of (2.11) is the system

$$(3.8) \quad x_a(t+1) = Ax_a(t) + Bu(t), \quad \phi_i(t) = C_i x_a(t) + E_i u(t),$$

where the matrices C_i and E_i are obtained from the matrix \bar{C}_i and \bar{E}_i by choosing the rows that correspond to the columns of $\bar{\phi}_i^T(t)$ that were selected to form $\phi_i^T(t)$. It might be pointed out that the state vector $x_a(t)$ and the matrices A and B are the same in all the associated-signal systems.

Now, recall that a linear system is called *output-reachable* if and only if every vector in its output space can be generated (reached) using a suitable input sequence. The following theorem relates the uniqueness of the solution of the design identity (2.6) to the output-reachability of the associated-signal systems.

THEOREM 3.1. *Assume that (2.6) is solvable. The solution is unique if and only if all r associated-signal systems of (2.11) are output-reachable.*

Proof of Theorem 3.1. The proof follows the lines of the proof of Theorem 4.1 for the scalar case presented in [1] and is omitted here. \square

Our original problem was to ensure that the estimation processes, which are based on (2.11), converge. It is known [2] that algorithms such as the RLS algorithm yield a sequence of estimates that converge exponentially fast to θ_i^* , provided that the sequence $\{\phi_i(t)\}$ of regression vectors is persistently exciting. The question is to find input sequences $\{u(t)\}$ such that all output sequences $\{\phi_i(t)\}$ of the (output-reachable) associated-signal systems will be persistently exciting. This problem is solved in the next section.

4. Persistent excitation of output-reachable MIMO plants. Consider the discrete-time, linear time-invariant plant

$$(4.1) \quad x_a(t+1) = Ax_a(t) + Bu(t), \quad y_a(t) = Cx_a(t) + Eu(t),$$

where $u(t) \in \mathbb{R}^m$, $y_a(t) \in \mathbb{R}^{p_a}$, and $x_a(t) \in \mathbb{R}^{n_a}$. In particular, this plant represents the associated-signal systems of § 3, with $y_a = \phi_i$, $n_a = m(l + \mu)$, and so on. System (4.1) can also be represented by the difference equation

$$(4.2) \quad \begin{aligned} & d_{n_a} y_a(t+1) + d_{n_a-1} y_a(t+2) + \cdots + d_1 y_a(t+n_a) + y_a(t+n_a+1) \\ &= G_{n_a} u(t+1) + G_{n_a-1} u(t+2) + \cdots + G_0 u(t+n_a+1), \end{aligned}$$

where d_i are the coefficients of a monic minimal polynomial for A , i.e., $a(z) = d_{n_a} + d_{n_a-1}z + \cdots + d_1 z^{n_a-1} + z^{n_a}$. The matrices $G_i \in \mathbb{R}^{p_a \times m}$ are defined by

$$(4.3) \quad G_i = \sum_{k=0}^i d_{i-k} M_k, \quad d_0 = 1,$$

where the matrices M_k are the Markov parameters, i.e.,

$$(4.4) \quad [M_0, \cdots, M_{n_a}] = [E, CB, CAB, \cdots, CA^{n_a-1}B].$$

We use the following definitions:

$$(4.5) \quad y_{a,j}(t) = [y_a(t+1), \cdots, y_a(t+j)],$$

$$(4.6) \quad \bar{u}_k(t) = [u^T(t+1), \cdots, u^T(t+k)]^T,$$

$$(4.7) \quad U_{k,j}(t) = [\bar{u}_k(t+1), \cdots, \bar{u}_k(t+j)],$$

$$(4.8) \quad G = [G_{n_a}, \cdots, G_0],$$

$$(4.9) \quad d = [d_{n_a}, \cdots, d_1, 1]^T.$$

Equation (4.2) can then be written as

$$(4.10) \quad G \cdot \bar{u}_{n_a+1}(t) = y_{a,n_a+1}(t) \cdot d.$$

DEFINITION. The sequence $\{y_a(t)\}$ is called *persistently exciting* if there exist $\varepsilon > 0$ and integers t_0 and N such that, for all integers $i \geq 0$,

$$(4.11) \quad \lambda_{\min}[y_{a,N}(t_0 + iN) y_{a,N}^T(t_0 + iN)] \geq \varepsilon > 0.$$

Since adaptation algorithms are known to be exponentially convergent, provided that the outputs of their associated-signal systems are persistently exciting, it is natural to find conditions on the inputs that result in this property. The following theorem addresses this issue.

THEOREM 4.1. Assume that plant (4.1) is output-reachable. If there exist $\varepsilon_1 > 0$ and integers t_1 and $N \geq n_a(m+1) + m$ such that, for all integers $i \geq 0$,

$$(4.12) \quad \lambda_{\min}[U_{n_a+1, N-n_a}(t_1 + iN) U_{n_a+1, N-n_a}^T(t_1 + iN)] \geq \varepsilon_1 > 0,$$

then the sequence $\{y_a(t)\}$ is persistently exciting for every initial state $x_a(0)$.

Remark 2. Note that (4.11) and (4.12) are equivalent to

$$(4.13) \quad \lambda_{\min} \left[\sum_{t=t_0+iN+1}^{t_0+iN+N} y_a(t) y_a^T(t) \right] \geq \varepsilon > 0$$

and

$$(4.14) \quad \lambda_{\min} \left[\sum_{t=t_1+iN+1}^{t_1+iN+N-n_a} \bar{u}_{n_a+1}(t) \bar{u}_{n_a+1}^T(t) \right] \geq \varepsilon_1 > 0.$$

Therefore, Theorem 4.1 shows that the persistency of excitation on y_a can be transformed into a similar condition on \bar{u}_{n_a+1} , which depends only on the input vector u . The

vector $\bar{u}_{n_a+1}(t)$ is obtained by stacking the vectors $u(t+1), \dots, u(t+n_a+1)$ on top of each other in a long vector. Since the dimension of \bar{u}_{n_a+1} is $(n_a+1)m$, the “span” of the sum $N-n_a$ must be greater than or equal to $(n_a+1)m$, and therefore the condition of Theorem 4.1 is obtained.

Proof of Theorem 4.1. Let $\alpha \in \mathbb{R}^{p_a}$ be a nonzero vector. Using (4.10),

$$\begin{aligned}
 \alpha^T G U_{n_a+1, N-n_a}(t_1) U_{n_a+1, N-n_a}^T(t_1) G^T \alpha &= \alpha^T G \left[\sum_{t=t_1+1}^{t_1+N-n_a} \bar{u}_{n_a+1}(t) \bar{u}_{n_a+1}^T(t) \right] G^T \alpha \\
 &= \left[\sum_{t=t_1+1}^{t_1+N-n_a} (\alpha^T y_{a, n_a+1}(t) d) (d^T y_{a, n_a+1}^T(t) \alpha) \right] \\
 &\leq \|d\|^2 \left[\sum_{t=t_1+1}^{t_1+N-n_a} \|\alpha^T y_{a, n_a+1}(t)\|^2 \right] \\
 (4.15) \quad &= \|d\|^2 \alpha^T \left[\sum_{t=t_1+1}^{t_1+N-n_a} y_{a, n_a+1}(t) y_{a, n_a+1}^T(t) \right] \alpha \\
 &\leq \|d\|^2 \alpha^T \left[\sum_{t=t_1+1}^{t_1+N-n_a} \sum_{j=t+1}^{t+n_a+1} y_a(j) y_a^T(j) \right] \alpha \\
 &\leq \|d\|^2 (n_a+1) \alpha^T \left[\sum_{j=t_1+2}^{t_1+N+1} y_a(j) y_a^T(j) \right] \alpha \\
 &\leq \|d\|^2 (n_a+1) \alpha^T [y_{a, N}(t_1+1) y_{a, N}^T(t_1+1)] \alpha.
 \end{aligned}$$

Since system (4.1) is output-reachable, the matrix G has full row rank, and, using (4.12),

$$\begin{aligned}
 \lambda_{\min}[y_{a, N}(t_1+1) y_{a, N}^T(t_1+1)] &\geq \frac{1}{\|d\|^2 (n_a+1)} \lambda_{\min}[G U_{n_a+1, N-n_a}(t_1) U_{n_a+1, N-n_a}^T(t_1) G^T] \\
 (4.16) \quad &\geq \frac{\lambda_{\min}(GG^T)}{\|d\|^2 (n_a+1)} \lambda_{\min}[U_{n_a+1, N-n_a}(t_1) U_{n_a+1, N-n_a}^T(t_1)] \\
 &\geq \varepsilon_1 \frac{\lambda_{\min}(GG^T)}{\|d\|^2 (n_a+1)} > 0.
 \end{aligned}$$

If we repeat the proof with $t_1 + iN$ instead of t_1 , and let $t_0 = t_1 + 1$,

$$(4.17) \quad \lambda_{\min}[y_{a, N}(t_0 + iN) y_{a, N}^T(t_0 + iN)] \geq \varepsilon_1 \frac{\lambda_{\min}(GG^T)}{\|d\|^2 (n_a+1)} = \varepsilon > 0,$$

and the proof is completed. \square

The following theorem shows that the results of Theorem 4.1 can be extended to cover situations where the input of the plant is calculated from an external reference input and state feedback. Input conditions are transferred to the reference input, assuming that the feedback gain matrix is held constant for sufficiently long periods between updates.

THEOREM 4.2. *Consider an output-reachable linear plant (4.1). Let the input sequence $u(t)$ be defined by the control law $u(t) = F_N(t)x_a(t) + v(t)$, where $v(t)$ is an external input and where $F_N(t)$ is a feedback gain matrix.*

If the matrix $F_N(t)$ is bounded and changes value only at times $t_i = t_0 + iN$, $i = 0, 1, 2, \dots$, with $N \geq n_a(m+1) + m$, and if the external input $v(t)$ satisfies the condition that

$$(4.18) \quad \lambda_{\min}[V_{n_a+1, N-n_a}(t_i) V_{n_a+1, N-n_a}^T(t_i)] \geq \varepsilon > 0,$$

then the sequence $\{y_a(t)\}$ is persistently exciting for every initial state $x_a(0)$.

Proof of Theorem 4.2. The proof is similar to the proof of Theorem 5.3 presented in [1] and is omitted here. \square

In the next sections, we show how Theorems 3.1 and 4.1 can be used to prove the global convergence of an adaptive control algorithm.

5. Adaptive pole placement for linear multivariable systems. It is assumed that the following parameters are known: $n, m, p, \mu_i, i = 1, 2, \dots, m$, and $\nu_i, i = 1, 2, \dots, p$. This is all the prior information required by the algorithm. The output of the plant (2.2) is given by

$$(5.1) \quad y(t) = R(D)P^{-1}(D)u(t).$$

The desired closed-loop dynamics are given by

$$(5.2) \quad y(t) = R(D)P^{*-1}(D)v(t),$$

where $v(t)$ is the external input and $P^*(D)$ is a polynomial matrix that characterizes the desired closed-loop pole locations. The control algorithm is an adaptive version of the control law

$$(5.3) \quad u(t) = Q^{-1}(D)[H(D)y(t) + K(D)u(t)] + v(t),$$

where $Q(D)$ is a fixed $(m \times m)$ polynomial matrix (all zeros of $\det(Q(D))$ are outside the unit circle) and where $H(D)$ and $K(D)$ are $(m \times p)$ and $(m \times m)$ controller matrices. The design equation of the controller is

$$(5.4) \quad H(D)R(D) + K(D)P(D) = Q(D)[P(D) - P^*(D)].$$

Since $R(D)$ and $P(D)$ are right coprime, there exist $(m \times p)$ and $(m \times m)$ matrices $J(D)$ and $I + S(D)$ that satisfy the Bezout identity

$$(5.5) \quad J(D)R(D) + [I + S(D)]P(D) = I.$$

Using (5.1), (5.4), and (5.5), we get the following nonminimal model of the plant:

$$(5.6) \quad \begin{aligned} & [H(D) + Q(D)P^*(D)J(D)]y(t) \\ & = [-K(D) - Q(D)P^*(D)S(D) + Q(D)(I - P^*(D))]u(t). \end{aligned}$$

This model is of the form (2.3). In this example, the general design identity (2.6) has the form

$$(5.7) \quad \begin{aligned} & [H(D) + Q(D)P^*(D)J(D)]R(D) \\ & = [-K(D) - Q(D)P^*(D)S(D) + Q(D)(I - P^*(D))]P(D). \end{aligned}$$

In the following theorem, we give conditions that ensure that (5.7) has a unique solution $\{H(D), K(D), J(D), S(D)\}$ such that $\{H(D), K(D)\}$ satisfy the design equation (5.4).

THEOREM 5.1. *Consider a plant of the form (2.1). Let $R(D)$ and $P(D)$ be the matrices in model (2.2). Let $Q(D)$ and $P^*(D)$ be $(m \times m)$ matrices of the form*

$$\begin{aligned} Q(D) &= \text{diag}[q_j(D)], \quad \deg[q_j(D)] = \nu + \mu - \mu_j, \quad q_j(0) = 1, \\ P^*(D) &= \text{diag}[p_j^*(D)], \quad \deg[p_j^*(D)] = \mu_j, \quad p_j^*(0) = 1, \end{aligned}$$

for $j = 1, 2, \dots, m$, where $q_j(D)$ and $p_j^*(D)$ are polynomials that have zeros outside the unit circle. Then (5.7) has a unique solution $\{H(D), K(D), J(D), S(D)\}$ of the form

$$(5.8) \quad \begin{aligned} H_i^j(D) &= \sum_{k=1+\nu+\mu-\mu_i-\nu_j}^{\nu+\mu-\mu_i} H_{ik}^j D^k, & K_i^j(D) &= \sum_{k=k_0(i,j)}^{\nu+\mu-\mu_i} K_{ik}^j D^k, \\ J_i^j(D) &= \sum_{k=1+\nu+\mu-\mu_i-\nu_j}^{\nu+\mu-\mu_i} J_{ik}^j D^k, & S_i^j(D) &= \sum_{k=k_0(i,j)}^{\nu+\mu-\mu_i} S_{ik}^j D^k, \end{aligned}$$

where $H_i^j(D)$ denotes the ij th element of $H(D)$, $H_{ik}^j, K_{ik}^j, J_{ik}^j, S_{ik}^j \in \mathbb{R}$, and where $k_0(i, j)$ is given by

$$(5.9) \quad k_0(i, j) = \begin{cases} 1 + \mu_j - u_i & \text{for } \mu_j \geq \mu_i, \\ 0 & \text{for } \mu_j < \mu_i. \end{cases}$$

The solution $\{H(D), K(D), J(D), S(D)\}$ is also the unique solution of (5.4), (5.5) under the conditions (5.8).

Remark 3. Note that any solution of (5.4), (5.5) is clearly a solution of (5.9). The reverse, however, is not so obvious. Theorem 5.1 shows that, under the degree constraints, (5.4) and (5.5) have unique solutions $\{H(D), K(D)\}$ and $\{J(D), S(D)\}$, which, together, constitute the unique solution of (5.7). To achieve this objective, we observe that constraints were imposed on the *lowest*, as well as the highest, degrees of $H(D)$, $K(D)$, $J(D)$, $S(D)$. Furthermore, compared to the scheme of [8], the degrees of the $q_j(D)$'s were increased from $\nu + \mu$ to $\nu + \mu - \mu_j$. This modification is not necessary for the uniqueness of the solutions of (5.4), (5.5), but was found necessary to prove that any solution of (5.7) is a solution of (5.4), (5.5).

Proof of Theorem 5.1. Preliminaries. The proof is easier to derive in terms of the forward shift operator, rather than in terms of the backward shift operator or delay D . In this framework, the proof is also similar to the proof for model reference adaptive control given in [4, p. 288]. We define

$$(5.10) \quad \begin{aligned} \bar{Q}(z) &= \text{diag} [z^{\nu+\mu-\mu_i}] Q(D) \big|_{D=z^{-1}}, & \bar{P}^*(z) &= P^*(D) \big|_{D=z^{-1}} \text{diag} [z^{\mu_j}], \\ \bar{H}(z) &= \text{diag} [z^{\nu+\mu-\mu_i}] H(D) \big|_{D=z^{-1}}, & \bar{K}(z) &= \text{diag} [z^{\nu+\mu-\mu_i}] K(D) \big|_{D=z^{-1}}, \\ \bar{J}(z) &= \text{diag} [z^{\nu+\mu-\mu_i}] J(D) \big|_{D=z^{-1}}, & \bar{S}(z) &= \text{diag} [z^{\nu+\mu-\mu_i}] S(D) \big|_{D=z^{-1}}, \\ \bar{R}(z) &= R(D) \big|_{D=z^{-1}} \text{diag} [z^{\mu_j}], & \bar{P}(z) &= P(D) \big|_{D=z^{-1}} \text{diag} [z^{\mu_j}]. \end{aligned}$$

Note that all these matrices are polynomial matrices in z . The constraints on the degrees of $H(D)$, $K(D)$, $J(D)$, and $S(D)$ in (5.8) may be shown to be *equivalent* to the following constraints on $\bar{H}(z)$, $\bar{K}(z)$, $\bar{J}(z)$, and $\bar{S}(z)$:

$$(5.11) \quad \partial_{c_j}(\bar{H}(z)) \leq \nu_j - 1, \quad \partial_{c_j}(\bar{J}(z)) \leq \nu_j - 1,$$

$$(5.12) \quad \partial_{r_i}(\bar{K}(z)) \leq \nu + \mu - \mu_i, \quad \partial_{c_j}(\bar{K}(z)) \leq \nu + \mu - \mu_j - 1,$$

$$(5.13) \quad \partial_{r_i}(\bar{S}(z)) \leq \nu + \mu - \mu_i, \quad \partial_{c_j}(\bar{S}(z)) \leq \nu + \mu - \mu_j - 1.$$

The constraints on $Q(D)$, $P^*(D)$ are equivalent to

$$(5.14) \quad \begin{aligned} \bar{Q}(z) &= \text{diag} [\bar{q}_j(z)], & \deg [\bar{q}_j(z)] &= \nu + \mu - \mu_j, & \bar{q}_j(0) &\neq 0, \\ \bar{P}^*(z) &= \text{diag} [\bar{p}_j^*(z)], & \deg [\bar{p}_j^*(z)] &= \mu_j, & \bar{p}_j^*(0) &\neq 0, \end{aligned}$$

provided that $\bar{q}_j(z)$, $\bar{p}_j(z)$ are *monic* polynomials with zeros inside the unit circle. With these definitions, (5.4) and (5.5) are equivalent to

$$(5.15) \quad \bar{H}(z)\bar{R}(z) + \bar{K}(z)\bar{P}(z) = \bar{Q}(z)(\bar{P}(z) - \bar{P}^*(z)),$$

$$(5.16) \quad \bar{J}(z)\bar{R}(z) + \bar{S}(z)\bar{P}(z) = z^{\nu+\mu}I - \text{diag} [z^{\nu+\mu-\mu_i}]\bar{P}(z),$$

while the design identity (5.7) is

$$(5.17) \quad \begin{aligned} (z^{\nu+\mu}\bar{H}(z) + \bar{Q}(z)\bar{P}^*(z)\bar{J}(z))\bar{R}(z) &= (-z^{\nu+\mu}\bar{K}(z) - \bar{Q}(z)\bar{P}^*(z)\bar{S}(z) \\ &\quad + z^{\nu+\mu}\bar{Q}(z) - \bar{Q}(z)\bar{P}^*(z) \\ &\quad \cdot \text{diag} [z^{\nu+\mu-\mu_j}])\bar{P}(z). \end{aligned}$$

From the properties of $R(D)$, $P(D)$, it follows that $\bar{R}(z)$, $\bar{P}(z)$ are right coprime, with $\partial_{c_j}(\bar{R}(z)) \leq \mu_j$, $\partial_{c_j}(\bar{P}(z)) = \mu_j$, and $\Gamma_c(\bar{P}(z)) = P(0)$ nonsingular (where $\Gamma_c(\bar{P}(z))$

denotes the matrix whose j th column contains the coefficients of z^{μ_j} in the j th column of $\bar{P}(z)$. It is a remarkable fact (cf. [10]) that there exists a *canonical* pair $(\bar{R}(z), \bar{P}(z))$, such that $\bar{P}(z)$ satisfies

$$(5.18) \quad \begin{aligned} \partial_{cj}(\bar{P}(z)) &= \mu_j, & [\Gamma_c(\bar{P}(z)) - I]_{ij} &= 0, & i \geq j, \\ \partial_{ri}(\bar{P}(z)) &= \mu_i, & \Gamma_r(\bar{P}(z)) - I &= 0. \end{aligned}$$

Note that the canonical $\bar{P}(z)$ is not only column-reduced, but also row-reduced. Similarly, there exists a *canonical left matrix fraction description* $(\tilde{P}(z), \tilde{R}(z))$ such that

$$(5.19) \quad \tilde{P}(z)\bar{R}(z) = \tilde{R}(z)\bar{P}(z)$$

and $\tilde{P}(z)$ satisfies

$$(5.20) \quad \begin{aligned} \partial_{ri}(\tilde{P}(z)) &= \nu_i, & [\Gamma_r(\tilde{P}(z)) - I]_{ij} &= 0, & j \geq i, \\ \partial_{cj}(\tilde{P}(z)) &= \nu_j, & \Gamma_c(\tilde{P}(z)) - I &= 0, \end{aligned}$$

with $\partial_{ri}(\tilde{R}(z)) \leq \nu_i$. With these preliminaries, we are ready to proceed with the proof of Theorem 5.1.

Existence. We first show that there exists a solution that satisfies (5.11). This result is available in the literature [11], but here we give a brief proof for completeness. Since $\bar{R}(z)$ and $\bar{P}(z)$ are right coprime, there exist matrices $\bar{U}(z)$ and $\bar{V}(z)$ such that

$$(5.21) \quad \bar{U}(z)\bar{R}(z) + \bar{V}(z)\bar{P}(z) = I.$$

The general solution of (5.15) is of the form

$$(5.22) \quad \begin{aligned} \bar{H}(z) &= \bar{Q}(z)(\bar{P}(z) - \bar{P}^*(z))\bar{U}(z) + \bar{Q}_1(z)\tilde{P}(z), \\ \bar{K}(z) &= \bar{Q}(z)(\bar{P}(z) - \bar{P}^*(z))\bar{V}(z) - \bar{Q}_1(z)\tilde{R}(z), \end{aligned}$$

and the solution of (5.16) is

$$(5.23) \quad \begin{aligned} \bar{J}(z) &= [z^{\nu+\mu}I - \text{diag}[z^{\nu+\mu-\mu_i}]\bar{P}(z)]\bar{U}(z) + \bar{Q}_2(z)\tilde{P}(z), \\ \bar{S}(z) &= [z^{\nu+\mu}I - \text{diag}[z^{\nu+\mu-\mu_i}]\bar{P}(z)]\bar{V}(z) - \bar{Q}_2(z)\tilde{R}(z), \end{aligned}$$

where $\bar{Q}_1(z)$ and $\bar{Q}_2(z)$ are arbitrary $(m \times p)$ polynomial matrices. From the polynomial matrix division theorem (cf. [9, p. 389], [4, p. 282]), there exist matrices $\bar{Q}_1(z)$ and $\bar{Q}_2(z)$ such that

$$(5.24) \quad \partial_{cj}(\bar{H}(z)) \leq \nu_j - 1, \quad \partial_{cj}(\bar{J}(z)) \leq \nu_j - 1.$$

It follows that $\bar{H}(z)$ and $\bar{J}(z)$ satisfy the degree constraints (5.11). Concerning the degree constraints on $\bar{K}(z)$, we multiply (5.15) on the left by $\text{diag}[z^{-(\nu+\mu-\mu_i)}]$ and on the right by $\text{diag}[z^{-\mu_j}]$ to obtain

$$(5.25) \quad \begin{aligned} &\text{diag}[z^{-(\nu+\mu-\mu_i)}]\bar{H}(z) \cdot \bar{R}(z) \text{diag}[z^{-\mu_j}] + \text{diag}[z^{-(\nu+\mu-\mu_i)}]\bar{K}(z) \cdot \bar{P}(z) \text{diag}[z^{-\mu_j}] \\ &= \text{diag}[z^{-(\nu+\mu-\mu_i)}]\bar{Q}(z) \cdot (\bar{P}(z) - \bar{P}^*(z)) \text{diag}[z^{-\mu_j}]. \end{aligned}$$

Since $\partial_{ri}(\bar{H}(z)) \leq \nu - 1 < \nu + \mu - \mu_i$, and using the properties of $\bar{R}(z)$, $\bar{P}(z)$, $\bar{P}^*(z)$, $\bar{Q}(z)$, it follows that

$$(5.26) \quad \lim_{z \rightarrow \infty} \text{diag}[z^{-(\nu+\mu-\mu_i)}]\bar{K}(z) = I - (\Gamma_c(\bar{P}(z)))^{-1} < \infty$$

and therefore

$$(5.27) \quad \partial_{ri}(\bar{K}(z)) \leq \nu + \mu - \mu_i.$$

The other constraint on $\bar{K}(z)$ is obtained by multiplying (5.15) on the left by $z^{-(\nu+\mu)}$, shown below:

$$(5.28) \quad \begin{aligned} & z^{-(\nu+\mu)} \cdot \bar{H}(z) \bar{R}(z) + \bar{K}(z) \operatorname{diag} [z^{-(\nu+\mu-\mu_j)}] \cdot \operatorname{diag} [z^{-\mu_i}] \bar{P}(z) \\ &= \operatorname{diag} [z^{-(\nu+\mu-\mu_i)}] \bar{Q}(z) \cdot \operatorname{diag} [z^{-\mu_i}] (\bar{P}(z) - \bar{P}^*(z)), \end{aligned}$$

where we used the fact that $\bar{Q}(z)$ is diagonal and that the product of diagonal matrices commutes. Since $\partial_{r_i}(\bar{H}(z) \bar{R}(z)) \leq \nu + \mu - 1$, it follows that

$$(5.29) \quad \lim_{z \rightarrow \infty} \bar{K}(z) \operatorname{diag} [z^{-(\nu+\mu-\mu_j)}] = I - (\Gamma_r(\bar{P}(z)))^{-1} = 0$$

and therefore

$$(5.30) \quad \partial_{c_j}(\bar{K}(z)) \leq \nu + \mu - \mu_j - 1.$$

The proof for the constraints on $\bar{J}(z)$ follows along identical lines.

Uniqueness. To prove uniqueness, we first establish that (5.17) can be satisfied only if (5.15) and (5.16) are satisfied (the converse being obvious). Rewrite (5.17) as

$$(5.31) \quad \begin{aligned} & z^{\nu+\mu} [\bar{H}(z) \bar{R}(z) + \bar{K}(z) \bar{P}(z) - \bar{Q}(z) (\bar{P}(z) - \bar{P}^*(z))] \\ &= -(\bar{Q}(z) \bar{P}^*(z)) [\bar{J}(z) \bar{R}(z) + \bar{S}(z) \bar{P}(z) - z^{\nu+\mu} I + \operatorname{diag} [z^{\nu+\mu-\mu_i}] \bar{P}(z)]. \end{aligned}$$

From the degree conditions and the properties of \bar{R}, \bar{P} , note that $\partial \bar{J}_{ik}(z) \leq \nu_k - 1 \leq \nu - 1$, $\partial \bar{R}_{kj}(z) \leq \mu_j \leq \mu$, $\partial \bar{S}_{ik}(z) \leq \nu + \mu - \mu_k - 1$, $\partial \bar{P}_{kj}(z) \leq \mu_k$. Furthermore, $\partial(z^{\nu+\mu-\mu_i} \bar{P}_{ik}(z) - z^{\nu+\mu}) \leq \nu + \mu - 1$. It follows that the maximal degree of any element in the right bracket in (5.31) is $\nu + \mu - 1$. However, the elements on the left side have $\nu + \mu$ zeros at $z = 0$, and $\bar{Q}(z) \bar{P}^*(z)$ is a diagonal matrix with elements that have no zeros at $z = 0$. Therefore, (5.31) can only be valid if both sides are equal to zero; i.e., if both (5.15) and (5.16) are satisfied.

Now, assume that there exists another solution $\bar{H}(z) + \delta \bar{H}(z)$, $\bar{K}(z) + \delta \bar{K}(z)$, $\bar{J}(z) + \delta \bar{J}(z)$, $\bar{S}(z) + \delta \bar{S}(z)$. It would then be necessary that the following homogeneous equations be satisfied:

$$(5.32) \quad \delta \bar{H}(z) \bar{R}(z) + \delta \bar{K}(z) \bar{P}(z) = 0, \quad \delta \bar{J}(z) \bar{R}(z) + \delta \bar{S}(z) \bar{P}(z) = 0.$$

Since $\bar{R}(z)$ and $\bar{P}(z)$ are coprime, and since $\partial_{c_j}(\delta \bar{H}(z)) \leq \nu_j - 1$ and $\partial_{c_j}(\delta \bar{J}(z)) \leq \nu_j - 1$, this implies (cf. [12]) that $\delta \bar{H}(z) = \delta \bar{K}(z) = \delta \bar{J}(z) = \delta \bar{S}(z) = 0$. \square

Expressions for the structured nonminimal model. We now show how the model (5.6) can be put in the form (2.3). The matrices in (5.6) can be written as follows:

$$(5.33) \quad \begin{aligned} H(D) &= \sum_{k=1}^{\nu+\mu-\mu_{\min}} H_k D^k, & K(D) &= \sum_{k=0}^{\nu+\mu-\mu_{\min}} K_k D^k, \\ J(D) &= \sum_{k=0}^{\nu+\mu-\mu_{\min}} J_k D^k, & S(D) &= \sum_{k=0}^{\nu+\mu-\mu_{\min}} S_k D^k, \end{aligned}$$

where $H_k, J_k \in \mathbb{R}^{m \times p}$, $K_k, S_k \in \mathbb{R}^{m \times m}$, and $\mu_{\min} = \min_{1 \leq j \leq m} \{\mu_j\}$. Let $\gamma = \nu + \mu - \mu_{\min}$. Substitution of (5.33) into (5.6) yields the following parameterization for (2.3):

$$(5.34) \quad C(D) = 0, \quad E(D) = Q(D)[I - P^*(D)],$$

$$(5.35) \quad B_0(D) = -I, \quad \beta_0 = K_0,$$

while, for $j = 1, 2, \dots, \gamma$,

$$(5.36) \quad A_j(D) = D^j \cdot I, \quad \alpha_j = H_j, \quad B_j(D) = -D^j \cdot I, \quad \beta_j = K_j,$$

and, for $j = \gamma + 1, \dots, 2\gamma + 1$,

$$(5.37) \quad \begin{aligned} A_j(D) &= Q(D)P^*(D)D^{j-\gamma-1}, & \alpha_j &= J_{j-\gamma-1}, \\ B_j(D) &= Q(D)P^*(D)D^{j-\gamma-1}, & \beta_j &= S_{j-\gamma-1}, \end{aligned}$$

with $m_a = m_b = 2\gamma + 1$, $l = \nu + \mu + \gamma = 2\nu + 2\mu - \mu_{\min}$, $r = m$.

The vectors $\bar{\theta}_i^*$, for $1 \leq i \leq m$, are given by

$$(5.38) \quad \bar{\theta}_i^* = [H_{i1}, \dots, H_{i\gamma}, J_{i0}, \dots, J_{i\gamma}, K_{i0}, \dots, K_{i\gamma}, S_{i0}, \dots, S_{i\gamma}],$$

$i = 1, 2, \dots, m,$

where H_{ik} , J_{ik} , K_{ik} , and S_{ik} are the i th row of the matrices H_k , J_k , K_k , and S_k (which are defined in (5.33)). The vectors $\bar{\phi}_i(t)$, $1 \leq i \leq m$ are given by

$$(5.39) \quad \begin{aligned} \bar{\phi}_i^T(t) &= [y^T(t-1), \dots, y^T(t-\gamma), q_i(D)p_i^*(D)y^T(t), \dots, q_i(D)p_i^*(D)y^T(t-\gamma), \\ &\quad u^T(t), \dots, u^T(t-\gamma), q_i(D)p_i^*(D)u^T(t), \dots, q_i(D)p_i^*(D)u^T(t-\gamma)], \end{aligned}$$

where $q_i(D)$ and $p_i^*(D)$ are the polynomials defined in Theorem 5.1. Each of the parameter vectors θ_i^* , $1 \leq i \leq m$ is obtained from $\bar{\theta}_i^*$ by deleting the elements that, according to the conditions in (5.8), are zero. In the same way, we obtain the vectors $\phi_i(t)$ from $\bar{\phi}_i(t)$, $1 \leq i \leq m$.

6. Stability and convergence properties. We now show that the general theorems of §§ 3 and 4 apply to the adaptive pole placement scheme. From Theorems 3.1 and 5.1, it follows that all m associated-signal systems are output-reachable, provided that the degree conditions in (5.8) are satisfied. By using Theorem 4.2, we obtain that all sequences $\{\phi_i(t)\}$, $1 \leq i \leq m$ (which are the associated-signal systems outputs) are persistently exciting, provided that the external input sequence $\{v(t)\}$ satisfies condition (4.18). We only need to show that the adaptive version of the control law (5.3) is of the form of the control law in Theorem 4.2.

The adaptive control law is given by

$$(6.1) \quad \begin{aligned} u(t) &= Q^{-1}(D)[H(D, t)y(t) + K(D, t)u(t)] + v(t) \\ &= H(D, t)y(t) + K(D, t)u(t) + Q(D)v(t) - (Q(D) - I)u(t), \end{aligned}$$

where $H(D, t)$ and $K(D, t)$ are the estimates of $H(D)$ and $K(D)$ at time t . For analysis purposes, it is useful to express (6.1) row by row, using (2.2) as follows:

$$(6.2) \quad \begin{aligned} u_i(t) &= [H_i(D, t)R(D) + K_i(D, t)P(D) - (q_i(D) - 1)P_i(D)]\xi(t) \\ &\quad + q_i(D)v_i(t), \quad i = 1, 2, \dots, m, \end{aligned}$$

where $u_i(t)$ and $v_i(t)$ are the i th component of $u(t)$ and $v(t)$; $H_i(D, t)$, $K_i(D, t)$, and $P_i(D)$ are the i th row of $H(D, t)$, $K(D, t)$, and $P(D)$; and where $q_i(D)$ are polynomials defined in Theorem 5.1. We will use the following definitions:

$$(6.3) \quad \begin{aligned} H_i(D, t)R(D) &= \sum_{k=1}^{\nu+2\mu-\mu_i} L_{ik}(t)D^k, \\ K_i(D, t)P(D) &= \sum_{k=1}^{\nu+2\mu-\mu_i} M_{ik}(t)D^k + K_i(0, t)P(D), \\ (q_i(D) - 1)P_i(D) &= \sum_{k=1}^{\nu+2\mu-\mu_i} N_{ik}D^k, \end{aligned}$$

where $L_{ik}(t)$, $M_{ik}(t)$, $N_{ik} \in \mathbb{R}^{m \times m}$. Substitution of (6.3) into (6.2) yields

$$\begin{aligned} u_i(t) &= \sum_{k=1}^{\nu+2\mu-\mu_i} [(L_{ik}(t) + M_{ik}(t) - N_{ik})D^k]\xi(t) + K_i(0, t)P(D)\xi(t) + q_i(D)v_i(t) \\ (6.4) \quad &= \sum_{k=1}^{\nu+2\mu-\mu_i} [(L_{ik}(t) + M_{ik}(t) - N_{ik})]\xi(t-k) + K_i(0, t)u(t) + q_i(D)v_i(t) \end{aligned}$$

for $i = 1, 2, \dots, m$. Equation (6.4) can be written as

$$(6.5) \quad u_i(t) = \bar{F}_i(t)x_a(t) + K_i(0, t)u(t) + q_i(D)v_i(t),$$

where $x_a(t)$ is the state vector of the associated-signal systems defined in (3.1). In this case, $l = 2\nu + 2\mu - \mu_{\min}$, $n_a = \dim[x_a] = m(l + \mu) = m(2\nu + 3\mu - \mu_{\min})$. $\bar{F}_i(t)$ is the following $(1 \times m(2\nu + 3\mu - \mu_{\min}))$ row vector:

$$\begin{aligned} (6.6) \quad \bar{F}_i(t) &= [L_{i1}(t) + M_{i1}(t) - N_{i1}, L_{i2}(t) + M_{i2}(t) - N_{i2}, \dots, L_{i\nu+2\mu-\mu_i}(t) \\ &\quad + M_{i\nu+2\mu-\mu_i}(t) - N_{i\nu+2\mu-\mu_i}, 0, \dots, 0]. \end{aligned}$$

By writing the m equations (6.5) for $i = 1, 2, \dots, m$, we obtain

$$(6.7) \quad u(t) = \bar{F}(t)x_a(t) + K(0, t)u(t) + Q(D)v(t),$$

where

$$(6.8) \quad \bar{F}(t) = [F_1^T(t), F_2^T(t), \dots, F_m^T(t)]^T.$$

From the constraints on $K_i^j(D)$ in (5.8), it follows that the matrix $[I - K(0, t)]$ is upper triangular and has a unit diagonal for all t . Therefore, its inverse always exists and is also upper triangular with unit diagonal. Hence, we can write

$$(6.9) \quad u(t) = F(t)x_a(t) + [I - K(0, t)]^{-1}Q(D)v(t),$$

where

$$(6.10) \quad F(t) = [I - K(0, t)]^{-1}\bar{F}(t).$$

In fact, it can be shown that $F(t)$ is given by

$$(6.11) \quad F(t) = [I - K(0, t)]^{-1} \cdot \left\{ \begin{bmatrix} \theta_1^T(t) & & & 0 \\ & \theta_2^T(t) & & \\ & & \ddots & \\ 0 & & & \theta_m^T(t) \end{bmatrix} \cdot T_1 - T_2 \right\},$$

where $\theta_i(t)$ are the parameter vectors estimates of θ_i^* (cf. (2.8)–(2.10) and equations following) and where T_1 and T_2 are *fixed* real matrices, which depend on the elements of $R(D)$, $P(D)$, and $Q(D)$.

Let $\{t_i\}$ be a sequence of integers such that $t_i = iN$, $i = 0, 1, 2, \dots$, where N is a positive integer to be determined later. The feedback gain matrix is held constant during each period of length N , and the adaptive control law is *modified* so that

$$(6.12) \quad u(t) = F_N(t)x_a(t) + w(t),$$

where

$$(6.13) \quad F_N(t) = F(t_i) \quad \text{for } t_i \leq t < t_{i+1},$$

$$(6.14) \quad w(t) = [I - K(0, t_i)]^{-1}\bar{v}(t) \quad \text{for } t_i \leq t < t_{i+1},$$

$$(6.15) \quad \bar{v}(t) = Q(D)v(t).$$

It is known (see [2]) that, with an RLS algorithm with covariance resetting, the estimates $\theta_i(t)$ remain in a bounded region of the parameter space. By the triangular property of $K(0, t_i)$, it follows that $(I - K(0, t_i))^{-1}$ is bounded. The matrices T_1 and T_2 are fixed so that, by (6.11), $F_N(t)$ remains bounded. The sequence $\{w(t)\}$ in (6.14) depends on the value of $K(0, t_i)$, but since $[I - K(0, t)]$ is bounded, it follows that if

$$(6.16) \quad \lambda_{\min}[\bar{V}_{n_a+1, N-n_a}(t_i) \bar{V}_{n_a+1, N-n_a}^T(t_i)] > \varepsilon_1 > 0,$$

then

$$(6.17) \quad \lambda_{\min}[W_{n_a+1, N-n_a}(t_i) W_{n_a+1, N-n_a}^T(t_i)] > \varepsilon > 0.$$

$Q(D)$ being fixed, the external input sequence $\{v(t)\}$ must be chosen so that (6.16) will be satisfied. Using Theorem 4.2, we obtain the following proposition that summarizes the results.

PROPOSITION 5.1. *Consider a linear minimal system (2.1). Assume that the observability indices ν_i , $1 \leq i \leq p$ and the controllability indices μ_i , $1 \leq i \leq m$ are known. Let $Q(D)$ and $P^*(D)$ be defined as in Theorem 5.1. Define m estimation equations of the form (2.11), for matrices $H(D)$, $K(D)$, $J(D)$, and $S(D)$ that satisfy the degree constraints in (5.8). Every parameter vector θ_i^* , $1 \leq i \leq m$ is estimated with an RLS algorithm with covariance resetting. The adaptive control law is given by*

$$(6.18) \quad u(t) = Q^{-1}(D)[H_N(D, t)y(t) + K_N(D, t)u(t) + Q(D)v(t)],$$

where $H_N(D, t)$ and $K_N(D, t)$ are the estimates of the matrices $H(D)$ and $K(D)$, updated periodically so that

$$H_N(D, t) = H(D, t_i), \quad K_N(D, t) = K(D, t_i) \quad \text{for } t_i \leq t < t_{i+1},$$

where $t_i = iN$, $i = 0, 1, 2, \dots$ and $N \geq m(2\nu + 3\mu - \mu_{\min})(m+1) + m$. The external input sequence $\{v(t)\}$ satisfies (6.16) (where $\bar{v}(t)$ is defined in (6.15)). Then the transfer matrix of the closed-loop system converges exponentially fast to $T_{cl}(D) = R(D)P^{*-1}(D)$, for every initial state of the system and for all initial conditions of the estimation algorithm.

7. Conclusions. In this paper, we showed how a multivariable adaptive pole placement algorithm could be designed so that parameter convergence is guaranteed under persistency of excitation conditions. More generally, it was proved that parameter convergence would follow, provided that a certain design identity had a unique solution, so that the results of this paper are applicable to a wide range of adaptive control algorithms.

An advantage of parameter convergence is that the closed-loop system asymptotically has the properties for which the controller was designed. In particular, the scheme presented here does not have the uncertainty of a matrix $U(D)$ found in [8] (present even with persistently exciting signals). On the other hand, more prior information is needed; that is, the observability indices must be known, in addition to the controllability indices. While the persistency of excitation conditions were used to assess stability, it is known that such conditions are not necessary to prove stability in adaptive control, but only to prove exponential convergence of the parameters to the nominal values (cf. [2], [4], [13], and [14] specifically for adaptive pole placement algorithms). The results of this paper may also be related to the work of [15], which discusses the minimum value of N in Theorem 4.2 for the SISO case, and to the work of [7], which transforms the persistency of excitation condition into a condition on the number of spectral components of the inputs (sufficient richness condition) in the case of multivariable identification. Special signals such that the persistency of excitation condition is satisfied were also investigated in [16].

Acknowledgments. The authors thank the anonymous reviewers for their constructive comments.

REFERENCES

- [1] M. HEYMANN, *Persistency of excitation results for structured nonminimal models*, IEEE Trans. Automat. Control, 33 (1988), pp. 112–116.
- [2] G. C. GOODWIN AND K. S. SIN, *Adaptive Filtering, Prediction and Control*, Prentice-Hall, Englewood Cliffs, NJ, 1984.
- [3] S. BOYD AND S. SASTRY, *Necessary and sufficient conditions for parameter convergence in adaptive control*, Automatica, 22 (1986), pp. 629–639.
- [4] S. SASTRY AND M. BODSON, *Adaptive Control: Stability, Convergence, and Robustness*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
- [5] B. D. O. ANDERSON AND R. M. JOHNSTONE, *Adaptive systems and time-varying plants*, Internat. J. Control, 37 (1983), pp. 367–377.
- [6] B. D. O. ANDERSON, *Adaptive systems, lack of persistency of excitation and bursting phenomena*, Automatica, 21 (1985), pp. 247–258.
- [7] M. DE MATHELIN AND M. BODSON, *Frequency domain conditions for parameter convergence in multivariable recursive identification*, Automatica, 6 (1990), pp. 757–767.
- [8] H. ELLIOTT, W. A. WOLOVICH, AND M. DAS, *Arbitrary adaptive pole placement for linear multivariable systems*, IEEE Trans. Automat. Control, 29 (1984), pp. 221–229.
- [9] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
- [10] S. BEGHELLI AND R. GUIDORZI, *A new input-output canonical form for multivariable systems*, IEEE Trans. Automat. Control, 21 (1976), pp. 692–696.
- [11] W. A. WOLOVICH AND P. J. ANTSAKLIS, *The canonical diophantine equations with applications*, SIAM J. Control Optim., 22 (1984), pp. 777–787.
- [12] G. D. FORNEY, *Minimal bases of rational vector spaces with applications to multivariable linear systems*, SIAM J. Control, 13 (1975), pp. 493–520.
- [13] J. W. POLDERMAN, *Adaptive Control & Identification: Conflict or Conflux?*, Ph.D. dissertation, University of Groningen, Groningen, the Netherlands, 1987.
- [14] R. LOZANO-LEAL AND G. C. GOODWIN, *A globally convergent adaptive pole placement algorithm without a persistency of excitation requirement*, IEEE Trans. Automat. Control, 30 (1985), pp. 795–798.
- [15] A. FEUER AND M. HEYMANN, *On minimum spanning blocks in discrete linear systems*, IEEE Trans. Automat. Control, 31 (1986), pp. 352–355.
- [16] Y. WILLNER, *Convergence of parameters in estimation and adaptive control of multivariable systems*, M.S. thesis (in Hebrew), Dept. of Electrical Engineering, Technion, Israel, 1986.
- [17] M. HEYMANN, J. H. LEWIS, AND G. MEYER, *Remarks on the adaptive control of linear plants with unknown high-frequency gain*, Systems Control Lett., 5 (1985), pp. 357–362.
- [18] P. R. KUMAR, *Convergence of adaptive control schemes using least-squares parameter estimates*, IEEE Trans. Automat. Control, 35 (1990), pp. 416–424.

OVERTAKING OPTIMAL REGULATION AND TRACKING OF PIECEWISE DIFFUSION LINEAR SYSTEMS*

ALAIN HAURIE† AND ARIE LEIZAROWITZ‡

Abstract. The infinite horizon optimal control of linear stochastic systems with quadratic cost integrand is studied, and the tracking of a periodic signal on an infinite time interval is considered. The system is exposed to three types of noises and is modeled by a nonhomogeneous linear stochastic control plant with modal and diffusion disturbances, where the dynamics switch at random times within a finite number of descriptions according to a Markov chain.

The *overtaking optimality* criterion is employed. Considering the expected cost, the existence of a unique optimal control is established for the above noisy control systems with *partial information*. This is realized as an affine feedback control of the best estimate of the state. Moreover, in the case of partial information, this feedback control is proved to be also *almost surely overtaking optimal*.

Key words. infinite horizon optimal control, the LQG problem, overtaking optimality, piecewise diffusion linear systems

AMS(MOS) subject classification. 93E20

1. Introduction. In this paper we study the infinite horizon optimal control of linear stochastic systems with quadratic cost integrand. In its most general form, the problem considered consists of tracking a periodic signal on an infinite time interval, where the system is modeled by a nonhomogeneous linear stochastic control plant with modal and diffusion disturbances. The simplest case considered corresponds to the regulator problem with the dynamics switching at random times within a finite set of descriptions, according to a Markov chain. In this latter case, it is quite easy to prove that under general assumptions of stabilizability and observability there exists a feedback control that minimizes the expected cost over the infinite horizon. For all the other cases, we must deal with cost expressions that diverge as the time interval increases indefinitely. An interesting way to deal with such unbounded cost criteria is furnished by the *overtaking optimality* concept. Initially introduced in the realm of economic growth models (see von Weizsäcker [25], Gale [11], Brock [4]), this concept has been used for more general deterministic control problems (Brock and Haurie [5], Carlson and Haurie [7], Leizarowitz [13]) and, more recently, for the analysis of regulation and tracking problems in the case of linear systems with stochastic diffusion disturbances (Leizarowitz [14]). This paper extends the analysis to the case where additional disturbances are represented by a jump Markov process and non-homogeneous offset terms. Moreover, for the piecewise diffusion regulator we establish almost sure overtaking optimality property of the optimal control, in the class of feedback controls.

The optimal control of stochastic systems with jump Markov disturbances has been studied by many authors. In Swarder [23] and Rishel [20], [21], a problem with a finite horizon is considered. A complete study of the infinite horizon case with a discounting factor has been developed by Vermes [24] and Davis [9]. A detailed analysis of a specific production planning model with a discounted convex cost

* Received by the editors June 18, 1990; accepted for publication (in revised form) February 15, 1991. This research has been supported by Natural Sciences and Engineering Research Council of Canada, Formation de chercheurs et action de recherche, Quebec, Actions Structurantes Ministère education supérieure et sciences, Quebec, and Fonds national de la recherche scientifique, Switzerland.

† Département d'économie commerciale et industrielle, Université de Genève, Geneva, Switzerland.

‡ Department of Mathematics, Technion-Israel Institute of Technology, 32000 Haifa, Israel.

integrand appears in Fleming, Sethi, and Soner [10]. We note that the exponential discounting term eliminates the divergence-of-costs difficulty.

A renewed interest in this class of systems stems from their use in the modeling of manufacturing flow control problems, in the contest of flexible manufacturing systems (FMS). In Olsder and Suri [19], Kimemia and Gershwin [12], Maimon and Gershwin [16], Akella and Kumar [1], Sharifnia [22], and Caramanis and Liberopoulos [6], an FMS is modeled as a linear system subject to jump random disturbances due to machine breakdowns and repairs. In these typical applications, there is no a priori known terminal time; therefore, infinite horizon considerations seem appropriate. However, the introduction of a discounting factor is not always appropriate since "infinity" in this context may correspond to a couple of months, if not weeks. In the above papers, we either deal rigorously with the discounted case (or the finite horizon case), or we use heuristic developments for the undiscounted case. The purpose of the present paper is to fill the gap by a rigorous treatment of the undiscounted case for linear quadratic systems.

The paper is organized as follows. In § 2 we describe the stochastic system and the three types of disturbances. We display some auxiliary results needed in the following sections, which are concerned with the simplest case where only the Markov jumps are present. Section 3 deals with the case where additional diffusion terms appear in the regulated system. The existence of a unique overtaking optimal control is established, and it is obtained in the well-known feedback control form.

We then turn to a more general case, where nonhomogeneous terms appear in the linear system and where, rather than being regulated, the system is supposed to track a given periodic trajectory. In § 4 we consider the infinite horizon Bellman equation for this problem. For this equation we construct a solution that is periodic in the time variable and quadratic in the space variable. We then interpret this function as the value function for a problem with a modified cost expression, where a time linear increasing function is subtracted from the original cost expression. This is described in § 5, where we also establish the existence of a unique overtaking optimal control given by the usual feedback control law.

In § 6 we consider the most general case, where, in addition to the three types of disturbances, there also appears uncertainty in the measurement of the state of the system. Thus there is an observed process that provides the information that is the basis for the choice of the control. We establish the existence of a unique overtaking optimal control, which is obtained by replacing the state variable with its best estimate in the feedback law described in § 5.

2. The framework, definitions, notations, and basic results. In this section we define the class of infinite time horizon, optimal stochastic control problems considered in this paper. We then give a solution to the simplest case, which consists of the regulation of a piecewise deterministic system.

2.1. A class of infinite horizon stochastic control problems. The system under consideration is

$$(2.1) \quad \begin{aligned} dx(t) &= [A_{j_t}x(t) + B_{j_t}u(t) + c_{j_t}] dt + G_{j_t} d\beta(t), \\ x(0) &= x_0, \quad j_0 = i, \quad x \in R^n, \quad u \in R^m, \end{aligned}$$

where $x(\cdot)$ is a stochastic process in R^n , $\beta(\cdot)$ is a p -dimensional Brownian motion, and $\{j_t\}_{t \geq 0}$ is a Markov process on the finite state space $\{1, \dots, N\}$ with generator $G = (g_{ij})_{i,j=1}^N$. The latter is such that any two states communicate on some finite time

interval. The matrices A_i , B_i , G_i and the vectors c_i , $1 \leq i \leq N$, are constant and of appropriate dimensions.

Let Y_t be the σ -algebra generated by $\{x(s), j_s, 0 \leq s \leq t\}$.

DEFINITION 2.1. An admissible control $u(\cdot)$ is a stochastic process defined on $[0, \infty)$ with values in R^m such that

- (i) $u(t)$ is Y_t -measurable for every $t \geq 0$,
- (ii) The solution of (2.1) corresponding to $u(\cdot)$, also called *the response* $x(\cdot)$ to $u(\cdot)$, is such that $t \rightarrow E|x(t)|^2$ is a bounded function on $[0, \infty)$.

Along with (2.1), a trajectory

$$(2.2) \quad \Gamma: [0, \infty) \rightarrow R^n$$

is given and assumed to be periodic with period T_0 , i.e., $\Gamma(t + T_0) = \Gamma(t)$ for all $t \geq 0$. The cost of using the control $u(\cdot)$ with its response $x(\cdot)$ over the $[0, T]$ interval is

$$(2.3) \quad C_T(u) = E_{i, x_0} \int_0^T [\|x(t) - \Gamma(t)\|_{Q_i}^2 + \|u(t)\|_{R_i}^2] dt,$$

where, for each $i = 1, \dots, N$, Q_i is a positive semidefinite symmetric matrix and R_i is a positive definite symmetric matrix. We employ the notation $\|x\|_Q^2 = x'Qx$, $\|u\|_R^2 = u'Ru$ for $x \in R^n$ and $u \in R^m$. The subscript (i, x_0) designates the initial values $j_0 = i$, $x(0) = x_0$.

As mentioned in the Introduction, except for one simple case that will be dealt with at the conclusion of this section, $C_T(u)$ diverges to infinity as T grows to infinity, for every admissible control $u(\cdot)$. This leads us to employ the overtaking optimality criterion, which is defined as follows.

DEFINITION 2.2. Let (i, x_0) be fixed initial values. The admissible control $u^*(\cdot)$ is *overtaking optimal* if, for every admissible control $u(\cdot)$,

$$(2.4) \quad \liminf_{T \rightarrow \infty} [C_T(u(\cdot)) - C_T(u^*(\cdot))] \geq 0.$$

Remark 2.3. If for some admissible $u(\cdot)$, the limit $\lim_{T \rightarrow \infty} C_T(u(\cdot))$ is finite, then the overtaking optimality notion coincides with ordinary optimality.

2.2. Regulation of linear piecewise deterministic systems. We consider the simplest case of linear piecewise deterministic systems

$$(2.5) \quad \begin{aligned} \dot{x}(t) &= A_{j_t}x(t) + B_{j_t}u(t), & x &\in R^n, \quad u \in R^m, \\ x(0) &= x_0, & j_0 &= i \end{aligned}$$

with the associated cost flow, which corresponds to $\Gamma(t) \equiv 0$.

$$(2.6) \quad C_T(u) = E_{i, x_0} \int_0^T [\|x(t)\|_{Q_{j_t}}^2 + \|u(t)\|_{R_{j_t}}^2] dt.$$

We are interested in "minimizing" $C_T(u)$ when T tends to infinity.

We assume here that (2.5) is stabilizable, namely, that there exist constant $m \times n$ matrices $\{F_j\}_{j=1}^N$ such that the feedback control $u(t) = F_{j_t}x(t)$ has a response $x(\cdot)$ that is stable in the quadratic mean sense

$$(2.7) \quad E_{i, x_0}|x(t)|^2 \rightarrow 0 \quad \text{as } t \rightarrow \infty.$$

In this situation the cost over the infinite horizon is finite, the minimization is in the usual sense, and there is no need to employ the overtaking concept. Although the result presented in this section may be obtained as a special case of the one described below for piecewise diffusion systems, we display it here for two reasons: First, we

will use it as a tool in subsequent sections and, second, we want to single out the piecewise deterministic situation as a separate case that has some interesting particular properties. Infinite horizon control of systems (2.5) and (2.6) and the relation with stabilization was studied by Morozan (see [17], [18]). Our developments are not the same as those in the work of Morozan and are given here for completeness. We assume that (2.5) is stabilizable, so that for a certain feedback control $u(t) = F_j x(t)$, the corresponding response $x(\cdot)$ satisfies (2.7). It follows from the linearity of (2.5) that the function $t \rightarrow E_{i, x_0} |x(t)|^2$ satisfies a linear differential equation (see [16, § 3]), which implies that the convergence in (2.7) is exponential. Then, given matrices $\{R_j\}_{j=1}^N$ and $\{Q_j\}_{j=1}^N$ as above, there exist controls u with finite cost $c_\infty(u)$. We thus may define

$$(2.8) \quad \phi_i(x_0) = \inf_{u(\cdot)} E_{i, x_0} \int_0^\infty [\|x(t)\|_{Q_i}^2 + \|u(t)\|_{R_i}^2] dt, \quad 1 \leq i \leq N,$$

which clearly satisfies $\phi_i(\lambda x) = \lambda^2 \phi_i(x)$ for every real λ . Moreover, it follows from the linearity of (2.5) and the fact that the integrand in (2.8) is quadratic that $\phi_i(x+y) + \phi_i(x-y) = 2\phi_i(x) + 2\phi_i(y)$ for every $x, y \in R^n$. It follows easily from these properties that the functions ϕ_i , $1 \leq i \leq N$, are quadratic and that there exist positive semidefinite matrices $\{K_j\}_{j=1}^N$ such that $\phi_j(x) = x' K_j x$, $1 \leq j \leq N$.

We claim that the following lemma holds.

LEMMA 2.4. Assume that for some k satisfying $1 \leq k \leq N$, the pair (A_k, Q_k) is observable. Then the matrices K_j , $1 \leq j \leq N$, are all positive definite.

Proof. It is enough to show that for every $1 \leq i \leq N$ and every $x_0 \neq 0$ we have $\phi_i(x_0) > 0$ (recall (2.8)), and for this it is enough to show that the integral in (2.8) is positive for every control (since the infimum is attained). Thus, by the positive definiteness of R_j , $1 \leq j \leq N$, we may consider only the zero control $u(t) \equiv 0$ for all $t \geq 0$. For this control let

$$p = P\{j_t = k, 1 \leq t \leq 2 | j_0 = i\};$$

then $p > 0$. Given an $x(0) = x_0$, $x_0 \neq 0$, there is a $\delta > 0$ such that for every sample path the response $x(\cdot)$ satisfies $|x(1)| \geq \delta$. The observability of (A_k, Q_k) implies that

$$\inf_{\substack{|x(1)| \geq \delta \\ v(\cdot)}} \int_1^2 [x(t)' Q_k x(t) + v(t)' R_k v(t)] dt \equiv \eta > 0,$$

where the infimum is over all measurable functions $v(\cdot)$ in $L_2([1, 2], R^m)$. It thus follows that for every $T > 2$, we have $C_T(u) \geq p\eta$, which proves that $\phi_i(x_0) > 0$ whenever $x_0 \neq 0$ and concludes the proof of the lemma. \square

We have the following characterization of stabilizable systems (2.5), which is linked to the infinite horizon optimization problem (2.5), (2.6).

PROPOSITION 2.5. Along with (2.5), consider sets of positive definite $m \times m$ matrices $\{R_j\}_{j=1}^N$ and positive semidefinite $n \times n$ matrices $\{Q_j\}_{j=1}^N$ such that at least one of the pairs (A_j, Q_j) , $1 \leq j \leq N$, is observable. A necessary and sufficient condition for the stabilization of (2.5) is that there exist positive definite matrices $\{K_j\}_{j=1}^N$ satisfying the Riccati system

$$(2.9) \quad A_i' K_i + K_i A_i - K_i B_i R_i^{-1} B_i' K_i + Q_i + \sum_{j=1}^N g_{ij} K_j = 0, \quad i = 1, \dots, N$$

for every $\{Q_j\}_{j=1}^N$ and $\{R_j\}_{j=1}^N$, as above. In this situation the feedback control $u^*(\cdot)$ given by

$$(2.10) \quad u^*(t) = -R_{j_t}^{-1} B_{j_t}' K_{j_t} x^*(t)$$

is stabilizing, and its response $x^*(\cdot)$ satisfies (2.7).

Proof. We first assume that (2.5) is stabilizable. Then, by Lemma 2.4 and the paragraph that precedes it, there are positive definite matrices K_i , $1 \leq i \leq N$, such that the $\phi_i(x)$ in (2.8) have the form $\phi_i(x) = x'K_ix$, $1 \leq i \leq N$.

The dynamic programming equation for system (2.5), (2.6) is

$$(2.11) \quad \min_u \left\{ x'Q_ix + u'R_iu + \sum_{j=1}^N g_{ij}x'K_jx + x'[A'_iK_i + K_iA_i]x + u'B'_iK_ix + x'K_iB_iu \right\} = 0, \\ i = 1, \dots, N,$$

where the unique minimum is attained by $u = -R_i^{-1}B'_iK_ix$. When this is substituted in (2.11), we find that $\{K_j\}_{j=1}^N$ satisfies (2.9). We will now show that the control (2.10) is stabilizing (and, in fact, is optimal for (2.5) and (2.6)). To this end, we consider the function

$$t \rightarrow E_{i,x_0}x(t)'K_{j_t}x(t).$$

It follows from the structure of the generator of $\{j_t\}_{t \geq 0}$ that

$$\begin{aligned} E_{i,x_0}x(t)'K_{j_t}x(t) &= x'_0K_ix_0 + E_{i,x_0} \int_0^t x(s)'[A'_{j_s}K_{j_s} + K_{j_s}A_{j_s}]x(s) ds \\ &\quad + E_{i,x_0} \int_0^t [u(s)'B'_{j_s}K_{j_s}x(s) + x(s)'K_{j_s}B_{j_s}u(s)]' ds \\ &\quad + E_{i,x_0} \int_0^t x(s)' \left(\sum_{l=1}^N g_{j_s,l}K_l \right) x(s) ds. \end{aligned}$$

Using the fact that $\{K_j\}_{j=1}^N$ satisfies (2.9), it follows that

$$\begin{aligned} E_{i,x_0}x(t)'K_{j_t}x(t) &= x'_0K_ix_0 - E_{i,x_0} \int_0^t [x(s)'Q_{j_s}x(s) + u(s)'R_{j_s}u(s)] ds \\ &\quad + E_{i,x_0} \int_0^t \|u(s) + R_{j_s}^{-1}B'_{j_s}K_{j_s}x(s)\|_{R_{j_s}}^2 ds, \end{aligned}$$

namely,

$$(2.12) \quad \begin{aligned} C_T(u) &= x'_0K_ix_0 - E_{i,x_0}x(T)'K_{j_T}x(T) \\ &\quad + E_{i,x_0} \int_0^T \|u(t) + R_{j_t}^{-1}B'_{j_t}K_{j_t}x(t)\|_{R_{j_t}}^2 dt. \end{aligned}$$

Thus, for the control u^* in (2.10), we obtain

$$C_T(u^*) = x'_0K_ix_0 - E_{i,x_0}x^*(T)'K_{j_T}x^*(T).$$

As remarked in the beginning of the proof, $x'_0K_ix_0$ is the infimal cost over $[0, \infty)$ starting at (x_0, i) . It hence follows that u^* is stabilizing and, in fact, is optimal since $c_\infty(u^*) = x'_0K_ix_0$.

To prove the other direction, we assume that for every given set $\{R_j\}_{j=1}^N$ and $\{Q_j\}_{j=1}^N$ as in the statement of the proposition, there are positive definite $\{K_j\}_{j=1}^N$ satisfying (2.9). We define the control u^* by (2.10), and it then follows from (2.12) that there exist controls with a finite cost on $[0, \infty)$. Proceeding as in the proof of the first part, this implies the existence of stabilizing controls. (In fact, it is easy to show that u^* is stabilizing and optimal.) \square

The stability of the response $x(\cdot)$ is, in fact, better than merely in quadratic mean.

PROPOSITION 2.6. *Let (2.5) be stabilizable by a feedback control (2.10) such that (2.7) holds. Then the response $x(\cdot)$ satisfies*

$$(2.13) \quad x(t) \rightarrow 0 \quad \text{a.s.}$$

Proof. We consider the functions

$$\phi_i(x, t) = E_{i,x}(t)' K_j x(t),$$

where $(x(\cdot), j)$ is the response to the control (2.10), which satisfies $j_0 = i$, $x(0) = x$. Then $\phi_i(x, t) = x' K_i(t) x$ for some positive definite matrices $\{K_i(t)\}_{i=1}^N$. It is easy to see that these matrices should satisfy a linear system

$$(2.14) \quad \frac{d}{dt} \begin{bmatrix} K_1(t) \\ \vdots \\ K_N(t) \end{bmatrix} = \mathbf{L} \begin{bmatrix} K_1(t) \\ \vdots \\ K_N(t) \end{bmatrix}, \quad \begin{bmatrix} K_1(0) \\ \vdots \\ K_N(0) \end{bmatrix} = \begin{bmatrix} K_1 \\ \vdots \\ K_N \end{bmatrix},$$

where \mathbf{L} is a linear operator from the space \mathbf{M} of N -tuple $n \times n$ matrices into itself, defined as follows: If $[M'_1, \dots, M'_N]' \in \mathbf{M}$, then

$$\mathbf{L} \begin{bmatrix} M_1 \\ \vdots \\ M_N \end{bmatrix}_i = M_i A_i + A_i' M_i + \sum_{j=1}^N g_{ij} M_j$$

(see, e.g., Morozan [17]). Since $K_i(t) \rightarrow 0$ as $t \rightarrow \infty$ for every $1 \leq i \leq N$, it follows from the linearity of (2.14) that the convergence is exponential. Thus there exist constants $c > 0$, $\alpha > 0$ such that $E|x(t)|^2 \leq c e^{-\alpha t}$ for all $t \geq 0$. It follows that for every $\varepsilon > 0$ we have $P(|x(t)| \geq \varepsilon) \leq (c/\varepsilon^2) e^{-\alpha t}$, which implies that $\sum_{k=0}^{\infty} P(|x(k)| \geq \varepsilon) < \infty$. Hence the Borel-Cantelli lemma implies that $x(k) \rightarrow 0$ almost surely as $k \rightarrow \infty$, so that (2.13) holds almost surely, and the proof is complete. \square

As indicated above, for a stabilizable linear piecewise deterministic system (2.5), (2.6), there exists a finite minimal cost over the infinite horizon. Thus the minimization is in the usual sense, and there is no need to invoke the overtaking concept. We summarize the discussion by stating the following result.

THEOREM 2.7. *Let (2.5) be stabilizable and consider the minimization of (2.6) with $T = \infty$. Assume that at least one of the pairs (A_j, Q_j) , $1 \leq j \leq N$, is observable. Then the feedback control u^* in (2.10) is the unique optimal control for the infinite horizon problem in the class of all nonanticipative controls.*

3. Linear piecewise diffusion stabilizable regulators. We now consider the system

$$(3.1) \quad \begin{aligned} dx(t) &= [A_{j_t} x(t) + B_{j_t} u(t)] dt + G_{j_t} dB(t), \\ x(0) &= x_0, \quad j_0 = i, \quad x \in R^n, \quad u \in R^m, \end{aligned}$$

with the associated cost flow (2.6). (Recall the paragraph that follows (2.1) and specifies the notations in this equation.)

Throughout this section, we assume that system (2.5) is stabilizable; thus there exist positive definite matrices $\{K_j\}_{j=1}^N$ satisfying (2.9). Recall Definition 2.1 of admissible controls.

Remark 3.1. Requirement (ii) in Definition 2.1 is quite natural in the context of stabilizable systems (2.5). If, e.g., the control is a stabilizing feedback control of the form $u(t) = \phi(x(t), j_t)$, where $\phi(\cdot, j)$ is Lipschitz continuous, then the Markov process $(x(t), j_t)_{t \geq 0}$ has an equilibrium measure $\{\nu_j(dx)\}_{j=1}^N$, and it is easy to see that unless $(1/T)C_T(u) \rightarrow \infty$, the function $t \rightarrow E|x(t)|^2$ converges to a finite limit.

For the forthcoming analysis we will need a version of Itô's lemma (see Åström [2]) applied to the following function:

$$\phi: R^n \times \{1, \dots, N\} \rightarrow R^1, \quad \phi(x, i) = x' K_i x.$$

The generator of the process $\{x(t), j_t\}_{t \geq 0}$ corresponding to a fixed constant control u is

$$(3.2) \quad (L^u f)(x, i) = \left(\frac{\partial f}{\partial x} \right)' (Ax + Bu) + \frac{1}{2} \sum_{j,k=1}^n a_{jk}^{(i)} \frac{\partial^2 f}{\partial x_j \partial x_k} (x, i) + \sum_{j=1}^N g_{ij} f(x, j),$$

where we denote $a_{jk}^{(i)} = (G_i' G_i)_{jk}$. Using the fact that $\{K_j\}_{j=1}^N$ satisfies (2.9), it follows from (3.2) that

$$(3.3) \quad \begin{aligned} x(T)' K_{j_T} x(T) - x_0' K_i x_0 = & \int_0^T \|u(t) + R_{j_t}^{-1} K_{j_t} x(t)\|_{R_{j_t}}^2 dt \\ & + \int_0^T \text{tr } G_{j_t}' K_{j_t} G_{j_t} dt + \int_0^T d\beta_t' G_{j_t}' K_{j_t} x(t) \\ & + \int_0^T x(t)' K_{j_t} G_{j_t} d\beta(t) - \int_0^T [\|x(t)\|_{Q_{j_t}}^2 + \|u(t)\|_{R_{j_t}}^2] dt. \end{aligned}$$

While taking expectation in (3.3), the stochastic integral terms drop (by virtue of (ii) of Definition 2.1) and we obtain

$$(3.4) \quad \begin{aligned} C_T(u) = & x_0' K_i x_0 + E_{i, x_0} \int_0^T \text{tr } G_{j_t}' K_{j_t} G_{j_t} dt \\ & + E_{i, x_0} \int_0^T \|u(t) + R_{j_t}^{-1} B_{j_t}' K_{j_t} x(t)\|_{R_{j_t}}^2 dt - E_{i, x_0} x(T)' K_{j_T} x(T). \end{aligned}$$

The main result of this section is the overtaking optimality of the control u^* in (2.10) in the class of all admissible controls. We consider the process

$$(3.5) \quad v(t) = u(t) + R_{j_t}^{-1} B_{j_t}' K_{j_t} x(t),$$

defined for an admissible control $u(\cdot)$ and its response $x(\cdot)$. If indeed the feedback law that determines $u^*(\cdot)$ is optimal, then $v(\cdot)$ measures the "deviation from optimality" of $u(\cdot)$. In particular, we have $v^*(t) \equiv 0$ for u^* . It follows from (3.4) that for every admissible control u

$$(3.6) \quad \begin{aligned} C_T(u) - C_T(u^*) = & E_{i, x_0} \int_0^T \|v(t)\|_{R_{j_t}}^2 dt + E_{i, x_0} x^*(T)' K_{j_T} x^*(T) \\ & - E_{i, x_0} x(T)' K_{j_T} x(T). \end{aligned}$$

If $u(\cdot)$ is such that

$$E_{i, x_0} \int_0^\infty \|v(t)\|_{R_{j_t}}^2 dt = \infty,$$

then $\lim_{T \rightarrow \infty} [C_T(u) - C_T(u^*)] = \infty$, and u^* overtakes u . We must thus consider only admissible controls $u(\cdot)$ for which $v(\cdot)$ satisfies

$$(3.7) \quad E_{i, x_0} \int_0^\infty \|v(t)\|_{R_{j_t}}^2 dt < \infty.$$

We will next prove that if (3.7) holds, then

$$\lim_{T \rightarrow \infty} E_{i, x_0} [x(T)' K_{j_T} x(T) - x^*(T)' K_{j_T} x^*(T)] = 0,$$

and it will follow from (3.6) that u^* overtakes u in this case, also.

The response $x^*(\cdot)$ to $u^*(\cdot)$ satisfies the equations

$$(3.8) \quad \begin{aligned} dx^*(t) &= [A_{j_t} - B_{j_t} R_{j_t}^{-1} B_{j_t}' K_{j_t}] x^*(t) dt + G_{j_t} d\beta(t), \\ x^*(0) &= x_0, \quad j_0 = i. \end{aligned}$$

Making use of the definition of $v(\cdot)$ in (3.5), it follows that the response $x(\cdot)$ to $u(\cdot)$ is a solution of

$$(3.9) \quad \begin{aligned} dx(t) &= \{[A_{j_t} - B_{j_t} R_{j_t}^{-1} B_{j_t}' K_{j_t}] x(t) + B_{j_t} v(t)\} dt + G_{j_t} d\beta(t), \\ x(0) &= x_0, \quad j_0 = i. \end{aligned}$$

Thus (3.8) and (3.9) imply that the process

$$(3.10) \quad y(t) = x(t) - x^*(t)$$

is almost surely absolutely continuous and satisfies the equation

$$(3.11) \quad \frac{dy}{dt} = [A_{j_t} - B_{j_t} R_{j_t}^{-1} B_{j_t}' K_{j_t}] y(t) + B_{j_t} v(t).$$

LEMMA 3.2. *Suppose that (2.5) is stabilizable, and let the matrices $\{K_i\}_{i=1}^N$ be as in (2.9). Suppose that at least one of the pairs (A_j, Q_j) , $1 \leq j \leq N$, is observable. Moreover, suppose that the control $u(\cdot)$ is such that (3.7) holds. Then*

$$(3.12) \quad \lim_{t \rightarrow \infty} E |y(t)|^2 = 0.$$

Proof. The solution $y(\cdot)$ of (3.11) is a solution of (2.5), corresponding to the admissible control

$$(3.13) \quad w(t) = -R_{j_t}^{-1} B_{j_t}' K_{j_t} y(t) + v(t).$$

For every $0 \leq T_0 < T < \infty$ we have

$$(3.14) \quad \begin{aligned} \int_{T_0}^T [\|y(t)\|_{Q_{j_t}}^2 + \|w(t)\|_{R_{j_t}}^2] dt &= \int_{T_0}^T y(t)' [Q_{j_t} + K_{j_t} B_{j_t} R_{j_t}^{-1} B_{j_t}' K_{j_t}] y(t) dt \\ &\quad - 2 \int_{T_0}^T y(t)' K_{j_t} B_{j_t} v(t) dt \\ &\quad + \int_{T_0}^T \|v(t)\|_{R_{j_t}}^2 dt. \end{aligned}$$

If (A_k, Q_k) is an observable pair and $T_0 \geq 1$, then for every sample path for which $j_t = k$ for every $T_0 \leq t \leq T_0 + 1$ we have

$$(3.15) \quad \int_{T_0}^{T_0+1} [\|y(t)\|_{Q_{j_t}}^2 + \|w(t)\|_{R_{j_t}}^2] dt \geq \varepsilon_1 |y(T_0)|^2.$$

If ε_2 is defined by

$$\varepsilon_2 = \inf_{T_0 \geq 1} \{P(j_t = k, T_0 \leq t \leq T_0 + 1 | j_0 = i)\},$$

then $\varepsilon_2 > 0$, and we obtain from (3.15) the uniform estimate

$$(3.16) \quad \begin{aligned} E_{i,x_0} \int_{T_0}^{T_0+1} [\|y(t)\|_{Q_{j_t}}^2 + \|w(t)\|_{R_{j_t}}^2] dt \\ \geq \varepsilon_3 E_{i,x_0} [|y(T_0)|^2 |j_t = k, T_0 \leq t \leq T_0 + 1] \end{aligned}$$

for some $\varepsilon_3 > 0$ and all $T_0 \geq 1$. Since for every $t > 0$, $y(t)$ is independent of $\{j_s\}_{s \geq t}$, the right-hand side of (3.16) is equal to $\varepsilon_3 E_{i,x_0} |y(T_0)|^2$. From the fact that $y(\cdot)$ is a solution of a linear differential equation with coefficients that are uniformly bounded for all the sample paths, it follows that $|y(t)| \leq \varepsilon_4 |y(T_0)|$ for some $\varepsilon_4 > 0$, for every $T_0 \geq 1$, $T_0 \leq t \leq T_0 + 1$, and every sample path. We thus obtain from (3.16) the following estimate:

$$(3.17) \quad E_{i,x_0} \int_{T_0}^T [\|y(t)\|_{Q_{j_t}}^2 + \|w(t)\|_{R_{j_t}}^2] dt \geq \alpha E_{i,x_0} \int_{T_0}^T |y(t)|^2 dt$$

for some $\alpha > 0$ and every $T_0 \geq 1$ and $T > T_0$ for which $T - T_0$ is an integer.

It follows from (3.11) that for every $T > 0$

$$\begin{aligned} E_{i,y_0} y(T)' K_{j_T} y(T) &= y_0' K_i y_0 + E_{i,y_0} \int_0^T \left[y(t)' K_{j_t} (A_{j_t} - B_{j_t} R_{j_t}^{-1} B_{j_t}' K_{j_t}) y(t) \right. \\ &\quad \left. + y(t)' (A_{j_t} - B_{j_t} R_{j_t}^{-1} B_{j_t}' K_{j_t})' K_{j_t} y(t) + y' \sum_{l=1}^N g_{j_t, l} K_l y \right] dt \\ &\quad + 2 E_{i,y_0} \int_0^T y(t)' K_{j_t} B_{j_t} v(t) dt, \end{aligned}$$

which by (2.9) can be written as

$$\begin{aligned} E_{i,y_0} y(T)' K_{j_T} y(T) &= y_0' K_i y_0 - E_{i,y_0} \int_0^T \left\{ y(t)' [Q_{j_t} + K_{j_t} B_{j_t} R_{j_t}^{-1} B_{j_t}' K_{j_t}] y(t) \right. \\ &\quad \left. - 2 y(t)' K_{j_t} B_{j_t} v(t) \right\} dt. \end{aligned}$$

By (3.14) and (3.17), using (3.7), this implies that

$$E_{i,y_0} |y(T)|^2 \leq a - \alpha E_{i,y_0} \int_0^T |y(t)|^2 dt$$

for some constants $a, \alpha > 0$ and for every $T > 0$. Then, clearly, $\int_0^T E_{i,y_0} |y(t)|^2 dt < \infty$, and since $(d/dt) E_{i,y_0} |y(t)|^2$ is bounded on $[0, \infty)$, this implies that $\lim_{t \rightarrow \infty} E_{i,y_0} |y(t)|^2 = 0$ as $t \rightarrow \infty$, concluding the proof of the lemma. \square

The main result of this section will now follow easily.

THEOREM 3.3. *Assume that (2.5) is stabilizable and at least one pair (A_j, Q_j) is observable. Then the feedback control u^* in (2.10) is overtaking optimal in the class of all admissible controls.*

Proof. As discussed in the paragraph following (3.6), u^* overtakes every admissible control $u(\cdot)$ for which $E \int_0^\infty |v(t)|^2 dt = \infty$, and it overtakes every admissible control $u(\cdot)$ for which

$$\lim_{T \rightarrow \infty} E_{i,x_0} [x(T)' K_{j_T} x(T) - x^*(T)' K_{j_T} x^*(T)] = 0$$

holds. However, the last equality follows from (3.7) and Lemma 3.2 (recalling (3.10)). \square

Remark. The quadratic function $\phi(x_0, i) = x' K_i x$ can now be given an interpretation of a minimal excess cost-to-go, namely, the limit of the cost left after subtracting

$$E_{i, x_0} \int_0^T \text{tr } G'_{j_t} K_{j_t} G_{j_t} dt.$$

This interpretation will be further developed in §§ 4 and 5 dealing with the more general periodic tracking problem.

In the rest of this section, we will show that u^* in (2.10) has an additional optimality property; namely, it is almost surely overtaking optimal in the class of all stationary controls.

DEFINITION 3.4. A stationary control is a feedback control $u(t) = \phi(x_j, j_t)$, where $\phi: R^n \times \{1, \dots, N\} \rightarrow R^m$ is Lipschitz continuous on R^n for every fixed $1 \leq j \leq N$, and u is such that $t \rightarrow E|x(t)|^2$ is a bounded function on $[0, \infty)$.

To establish almost surely overtaking optimality we will need the following assumption.

Assumption A. Let $\tilde{A}_i = A_i - B_i R_i^{-1} B_i' K_i$ for $1 \leq i \leq N$ and let $L(x)$ be the Lie algebra generated by $\{\tilde{A}_1, x, \dots, \tilde{A}_N, x\}$. Then $L(x)$ has rank n for every $x \neq 0$.

THEOREM 3.5. Assume that (2.5) is stabilizable, that at least one of the pairs (A_j, Q_j) $1 \leq j \leq N$, is observable, and that Assumption A holds. Then the control u^* is almost surely overtaking optimal in the class of stationary controls; namely, for every such control u there exists a positive real-valued random variable $\omega \rightarrow T(\omega)$ such that almost surely

$$(3.18) \quad \int_0^T [\|x^*(t)\|_{Q_i}^2 + \|u^*(t)\|_{R_i}^2] dt < \int_0^T [\|x(t)\|_{Q_i}^2 + \|u(t)\|_{R_i}^2] dt$$

for all $T > T(\omega)$.

Proof. With every stationary control $u(\cdot)$, there is associated a Markov process $\{x(t), j_t\}_{t \geq 0}$, where $x(\cdot)$ is the response to $u(\cdot)$. It is enough to consider stationary controls for which $\{x(t), j_t\}_{t \geq 0}$ is positively recurrent, since other controls have infinite cost growth rate. Then by the ergodic theorem

$$(3.19) \quad \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \|u(t) + R_{j_t}^{-1} K_{j_t} B_{j_t}' x(t)\|_{R_{j_t}}^2 dt = \sum_{j=1}^N \int_{R^n} \|\phi(x, j) + R_j^{-1} K_j B_j' x\|_{R_j}^2 \nu_j(dx),$$

where $\{\nu_j(dx)\}_{j=1}^N$ is an equilibrium measure of $\{x(t), j_t\}_{t \geq 0}$. The right-hand side of (3.19) is strictly positive unless $u = u^*$. If, however, $u = u^*$ on the support of $\{\nu_j(dx)\}_{j=1}^N$ then, by Assumption A, the support of $\{\nu_j(dx)\}_{j=1}^N$ is $R^n \times \{1, \dots, N\}$. It thus follows from (3.3) that to prove the theorem, it is sufficient to show that

$$(3.20) \quad \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x(t)' K_{j_t} G_{j_t} d\beta_t = 0 \quad \text{a.s.}$$

(since then the limit $\lim_{T \rightarrow \infty} (1/T)x(T)' K_{j_T} x(T)$ exists almost surely and must be equal to zero. It then follows that for every $u \neq u^*$ we have that almost surely

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T [\|x^*\|_{Q_i}^2 + \|u^*\|_{R_i}^2] dt < \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T [\|x(t)\|_{Q_i}^2 + \|u(t)\|_{R_i}^2] dt,$$

which implies (3.18)). To prove (3.20) we consider the martingale

$$M_T = \int_0^T x(t)' K_{j_t} G_{j_t} d\beta_t, \quad 0 \leq T < \infty,$$

and it will be enough to prove that $(1/k)M_k \rightarrow 0$ almost surely as $k \rightarrow \infty$. Since $E|x(t)|^2 \leq C_1$ for some constant $C_1 > 0$ and all $t \geq 0$, it follows that $E|(1/T)M_T|^2 \leq (C_2/T^2)$ for

some $C_2 > 0$ and all $T > 0$. We have then, by Chebyshev's inequality and the Borel-Cantelli lemma, that

$$(3.21) \quad \frac{1}{k^2} M_{k^2} \rightarrow 0 \quad \text{a.s. as } k \rightarrow \infty.$$

The following standard argument (see Chung [8, Thm. 4.12, p. 103]) extends the validity of (3.21) to all the integers rather than merely for the subsequence of squares. Every integer j can be written as $j = k^2 + p$, where $0 \leq p \leq 2k$, and we have, for the above constant C_2 , the estimate $E[M_j - M_{k^2}]^2 \leq pC_2$. It then follows that

$$\sum_{k=1}^{\infty} \sum_{j=k^2}^{(k+1)^2-1} E \left(\frac{M_j - M_{k^2}}{j} \right)^2 \leq C_2 \sum_{k=1}^{\infty} \frac{1}{k^4} \left(\sum_{p=1}^{2k} p \right) < \infty,$$

which implies that

$$(3.22) \quad \frac{M_j - M_{k^2}}{j} \rightarrow 0 \quad \text{a.s. as } k \rightarrow \infty \text{ and } j \text{ satisfies } k^2 \leq j < (k+1)^2.$$

Since

$$\frac{M_j}{j} = \left[\frac{M_{k^2}}{k^2} + \frac{M_j - M_{k^2}}{j} \left(1 + \frac{p}{k^2} \right) \right] \frac{k^2}{j}$$

and $p/k^2 \rightarrow 0$, $k^2/j \rightarrow 1$ as $k \rightarrow \infty$, it follows from (3.21) and (3.22) that $M_j/j \rightarrow 0$ almost surely as $j \rightarrow \infty$, which implies (3.20). As explained in the beginning, this concludes the proof of the theorem. \square

4. The infinite horizon Bellman equation. In this section we consider a functional equation, called the infinite horizon Bellman equation, and construct a solution to it. This equation plays a central role in our study of tracking a periodic signal with piecewise diffusion systems that are disturbed by modal jumps and inhomogeneous offset terms, developed in § 5.

To motivate our study of this equation, the following heuristic discussion is proposed for the case of piecewise deterministic systems. Due to the nonzero tracked signal and the Markov jumps of the system, we expect the cost to grow at some minimal rate μ , but still expect the expressions $C_T(u) - \mu T$ to remain bounded as $T \rightarrow \infty$, for the better controls u . Thus we write the Bellman equation for the excess cost, namely, the cost left after subtracting the linear part μT from the cost (2.6). Moreover, since the problem is defined on infinite time interval with a periodic signal to be tracked, the excess cost function should also be periodic.

The infinite horizon Bellman equation for the linear piecewise-deterministic systems is thus

$$(4.1) \quad \left(\frac{\partial \phi}{\partial t} \right) (i, x, t) + \min_u \left\{ \frac{1}{2} \|x - \Gamma(t)\|_{Q_i}^2 + \frac{1}{2} \|u\|_{R_i}^2 + \left(\frac{\partial \phi}{\partial x} \right)' (i, x, t) \cdot [A_i x + B_i u + c_i] \right. \\ \left. + \sum_{j=1}^N g_{ij} \phi(j, x, t) \right\} = \mu, \quad 1 \leq i \leq N,$$

which is an equation both for the function $(i, x, t) \rightarrow \phi(i, x, t)$ and the constant μ .

In the case of piecewise diffusion systems, we are led to consider the following Bellman equation:

$$(4.1') \quad \left(\frac{\partial \phi}{\partial t} \right) (i, x, t) + \min \{ \dots \} + \frac{1}{2} \sum_{j,k=1}^n a_{jk}^{(i)} \left(\frac{\partial^2 \phi}{\partial x_j \partial x_k} \right) (i, x, t) = \mu_i,$$

$$1 \leq i \leq N.$$

We try a solution of (4.1) or (4.1'), which is of the form

$$(4.2) \quad \phi(i, x, t) = \frac{1}{2}x'K_ix + q_i(t)' \cdot x + v_i(t),$$

where the functions $q_i(\cdot)$ and $v_i(\cdot)$ are periodic of period T_0 , which is the period of $\Gamma(\cdot)$.

Note that, with this form of solution, the second partial derivatives are constants, and so we could rewrite (4.1') as

$$\left(\frac{\partial \phi}{\partial t}\right)(i, x, t) + \min\{\dots\} + \frac{1}{2} \sum_{j,k=1}^n a_{jk}^{(i)} \left(\frac{\partial^2 \phi}{\partial x_j \partial x_k}\right) = \mu + \frac{1}{2} \sum_{j,k=1}^n a_{jk}^{(i)} \frac{\partial^2 \phi}{\partial x_j \partial x_k}$$

so that, a solution to (4.1') would also provide a solution to (4.1). The minimum in (4.1) is attained at

$$(4.3) \quad u = -R_i^{-1}B_i'[K_ix + q_i(t)],$$

so that (4.1) is

$$(4.4) \quad \left(\frac{dq_i}{dt}\right)' \cdot x + \frac{dv_i}{dt} + \frac{1}{2} \|x - \Gamma(t)\|_{Q_i}^2 - \frac{1}{2} (K_ix + q_i)' B_i R_i^{-1} B_i' (K_ix + q_i) \\ + (K_ix + q_i)' (A_ix + c_i) + \sum_{j=1}^N g_{ij} \left[\frac{1}{2} x' K_j x + q_j' \cdot x + v_j \right] = \mu.$$

The quadratic term in (4.4) vanishes if the matrices K_i , $1 \leq i \leq N$, satisfy the system of matrix algebraic Riccati equations

$$(4.5) \quad K_i A_i + A_i' K_i - K_i B_i R_i^{-1} B_i' K_i + \sum_{j=1}^N g_{ij} K_j + Q_i = 0, \quad i = 1, \dots, N.$$

Collecting the linear terms and equating to zero yields

$$(4.6) \quad \frac{dq_i}{dt} - Q_i \Gamma(t) + (A_i' - K_i B_i R_i^{-1} B_i') q_i + K_i c_i + \sum_{j=1}^N g_{ij} q_j = 0, \quad i = 1, \dots, N.$$

We look for a solution $\{q_1(\cdot), \dots, q_N(\cdot)\}$ to (4.6), which is periodic with period T_0 . If $\Phi(t, t_0)$ is the fundamental solution of the corresponding homogeneous equation

$$(4.7) \quad \frac{dq_i}{dt} + (A_i' - K_i B_i R_i^{-1} B_i') q_i + \sum_{j=1}^N g_{ij} q_j = 0, \quad i = 1, \dots, N,$$

then the solution $q(\cdot)$ of (4.6) satisfies $q(T_0) = \Phi(T_0, 0)q(0) + \rho(T_0)$, where $\rho(T_0)$ depends on $\Gamma(\cdot)$ and c . A periodic solution is obtained if $q(T_0) = q(0)$, and such a unique $q(0)$ is guaranteed if $I - \Phi(T_0, 0)$ is nonsingular, which holds if (4.7) does not have a nontrivial periodic solution of period T_0 . This property of (4.7) is, however, generic in a sense that will be precisely described next.

Our plant is determined by the collection of matrices

$$\{A_1, A_2, \dots, A_N, B_1, B_2, \dots, B_N, G\}.$$

Denote such a collection by S and let S be the metric space of all possible collections S with the following metric:

$$\rho(S, S') = \sum_{i=1}^N \|A_i - A_i'\| + \sum_{i=1}^N \|B_i - B_i'\| + \|G - G'\|,$$

where

$$S = \{A_1, A_2, \dots, A_N, B_1, B_2, \dots, B_N, G\}$$

and

$$S' = \{A'_1, A'_2, \dots, A'_N, B'_1, B'_2, \dots, B'_N, G'\}.$$

We say that a certain property is *generic for systems in S* if the set of systems that possess this property is open and dense in S.

Throughout the rest of the paper we will assume the following statement.

Assumption B.

- (i) At least one of the pairs (A_i, Q_i) , $1 \leq i \leq N$, is observable.
- (ii) Let $T_0 > 0$ be a given period time and $G = (g_{ij})_{i,j=1}^N$ a generator of a Markov process $\{j_t\}_{t \geq 0}$ such that any two states communicate on a finite time interval.
- (iii) Equation (4.7) does not have a nontrivial periodic solution of period T_0 .

PROPOSITION 4.1. *Let $T_0 > 0$ be given. Then the set of systems for which Assumption B holds is generic in S.*

Proof. If we consider the vectors (q_1, \dots, q_N) in (4.7) as a vector Q in R^{Nn} , then $Q(\cdot)$ satisfies a linear differential equation

$$(4.8) \quad \frac{dQ(t)}{dt} = MQ(t),$$

where M is an $(Nn \times Nn)$ -dimensional matrix that is related to the matrices coefficients in (4.7) in an obvious way. In particular, the entries of M are linear functions of the variables g_{ij} , $1 \leq i, j \leq N$. There is a nontrivial periodic solution to (4.8) for period T_0 if and only if unity is an eigenvalue of the matrix e^{MT_0} . From this, it clearly follows that the set of systems for which Assumption B holds is open in S.

To prove that this set is also dense, let S be a point in S, say $S = \{A_1, A_2, \dots, A_N, B_1, B_2, \dots, B_N, G\}$ and for every $\gamma > 0$ let $S_\gamma = \{A_1, A_2, \dots, A_N, B_1, B_2, \dots, B_N, \gamma G\}$. Let the matrix M_γ correspond to S_γ in the same way that the matrix M in (4.8) corresponds to S . Then the matrix function $\gamma \rightarrow \exp(M_\gamma T_0)$ is analytic (since the entries of M are linear functions of the entries of G). Therefore the scalar function $\gamma \rightarrow \det[I - \exp(M_\gamma T_0)]$ is analytic and has isolated zeros. It follows that for some $\varepsilon > 0$, every $\gamma \neq 1$ that satisfies $1 - \varepsilon < \gamma < 1 + \varepsilon$ is such that $\det[I - \exp(M_\gamma T_0)] \neq 0$, proving that S_γ satisfies Assumption B. Since γ can be chosen arbitrarily close to 1, this proves that the set of systems for which Assumption B holds is dense in S, and concludes the proof of the proposition. \square

For the free term in (4.4), we obtain

$$(4.9) \quad \frac{dv_i}{dt} + \sum_{j=1}^N g_{ij} v_j + \frac{1}{2} \|\Gamma(t)\|_{Q_i}^2 - \frac{1}{2} q_i(t)' B_i R_i^{-1} B_i' q_i(t) + q_i(t)' \cdot c_i = \mu, \quad i = 1, \dots, N,$$

which is of the form

$$(4.10) \quad \frac{dv_i}{dt} + \sum_{j=1}^N g_{ij} v_j = \mu - \rho_i(t), \quad i = 1, \dots, N,$$

where the functions $\rho_i(\cdot)$ are all periodic of period T_0 . From the explicit solution

$$v(t) = e^{-Gt} v(0) + \int_0^t e^{-G(t-s)} \begin{bmatrix} \mu - \rho_1(s) \\ \vdots \\ \mu - \rho_N(s) \end{bmatrix} ds,$$

it follows that the periodicity condition $v(0) = v(T_0)$ may be expressed by the requirement that

$$(4.11) \quad (I - e^{GT_0})v(0) = \beta - \mu\alpha,$$

where

$$(4.12) \quad \alpha = \int_0^{T_0} e^{Gt} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} dt, \quad \beta = \int_0^{T_0} e^{Gt} \begin{bmatrix} \rho_1(t) \\ \vdots \\ \rho_N(t) \end{bmatrix} dt.$$

We look for a $\nu(0) \in R^N$ and a scalar μ for which (4.11) holds. Since the matrix e^{GT_0} is a transition probability matrix, it follows from the property of G as mentioned in Assumption B that unity is a simple eigenvalue of e^{GT_0} . Hence the subspace of R^N

$$(4.13) \quad Y = \text{Im} [I - e^{GT_0}]$$

is $N - 1$ -dimensional. As long as α is not contained in Y , there is for every $\beta \in R^N$ a unique scalar μ for which (4.11) holds.

PROPOSITION 4.2. *Let α and Y be as in (4.12) and (4.13), respectively. Then the following holds:*

$$(4.14) \quad \alpha \notin Y.$$

Proof. Since G is a generator, it follows that

$$G \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = 0;$$

hence α in (4.11) is given by

$$\alpha = T_0 \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}.$$

If $\alpha \in Y$ then for some $v \in R^N$ the following equality holds:

$$e^{GT_0}v = v + T_0 \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix},$$

and iterating this relation we obtain that

$$(4.15) \quad e^{kT_0G}v = v + kT_0 \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

for every $k \geq 1$. However, since e^{GT_0} is a transition probability matrix we have, for every $v \in R^N$,

$$\max_{1 \leq i \leq N} |(e^{kT_0G}v)_i| \leq \max_{1 \leq i \leq N} |v_i|,$$

which clearly contradicts (4.15) as $k \rightarrow \infty$. This contradiction proves (4.14) and concludes the proof of the proposition. \square

To summarize the discussion in this section, we have constructed a solution to the infinite horizon Bellman equation (4.1), which is of the form (4.2). The matrices K_i , $1 \leq i \leq N$, are the solutions to (4.5). The functions $q_i(\cdot)$, $1 \leq i \leq N$ are the unique solution to (4.6) if (4.7) does not have a nontrivial periodic solution of period T_0 , which, by Proposition 4.2, is the generic case. The constant μ is uniquely determined by (4.11), (4.12), and $\nu(\cdot)$ is a solution to (4.9).

5. Tracking a periodic signal with piecewise diffusion systems. We consider system (2.1) with the cost expression (2.3), to which corresponds the Bellman equation (4.1') with the associated solution (4.2), as described in § 4. From Itô's lemma applied to the function ϕ in (4.2), we obtain

$$(5.1) \quad \begin{aligned} & \frac{1}{2}x'(T)'K_{j_T}x(T) + q_{j_T}(T)' \cdot x(T) + \nu_{j_T}(T) = \frac{1}{2}x'_0K_ix_0 + q_i(0)' \cdot x_0 + \nu_i(0) \\ & + \int_0^T \left\{ \frac{d\nu_{j_t}}{dt} + x(t)' \cdot \frac{dq_{j_t}}{dt} + [A_{j_t}x + B_{j_t}u + c_{j_t}]'(K_{j_t}x(t) + q_{j_t}) \right. \\ & \left. + \sum_{j=1}^N g_{j_t,j}\phi(j, x, t) \right\} dt + \int_0^T d\beta'_i G'_{j_t} [K_{j_t}x(t) + q_{j_t}] + \int_0^T \text{tr } G'_{j_t} K_{j_t} G_{j_t} dt. \end{aligned}$$

Simple computations using relations (4.5), (4.6), and (4.9) yield that the integrand in the first term of (5.1) is equal to

$$(5.2) \quad \frac{1}{2}\|u + R_{j_t}^{-1}B'_{j_t}[K_{j_t}x + q_{j_t}]\|_{R_{j_t}}^2 - \frac{1}{2}u'R_{j_t}u - \frac{1}{2}\|x - \Gamma(t)\|_{Q_{j_t}}^2 + \mu.$$

For example, the quadratic term in the integrand in (5.1) is

$$x' \left[\frac{1}{2} (A'_{j_t} K_{j_t} + K_{j_t} A_{j_t}) + \frac{1}{2} \sum_{j=1}^N g_{j_t,j} K_j \right] x,$$

while the quadratic term in (5.2) is

$$x' [\frac{1}{2} K_{j_t} B_{j_t} R_{j_t}^{-1} B'_{j_t} K_{j_t} - \frac{1}{2} Q_{j_t}] x,$$

and these are equal by (4.5). The linear term in u in both expressions is $u'B'_{j_t}[K_{j_t}x + q_{j_t}]$, and it is easy to verify that the linear terms in x , as well as the free terms in the two expressions, coincide. It thus follows from (5.1) that the cost process $\{c_T(u)\}_{T \geq 0}$ is given by

$$(5.3) \quad \begin{aligned} c_T(u) = & \int_0^T \mu_{j_t} dt + \int_0^T \text{tr } G'_{j_t} K_{j_t} G_{j_t} dt \\ & + \frac{1}{2} \int_0^T \|u(t) + R_{j_t}^{-1}B'_{j_t}[K_{j_t}x(t) + q_{j_t}]\|_{R_{j_t}}^2 dt \\ & + \int_0^T d\beta'_i G'_{j_t} [K_{j_t}x(t) + q_{j_t}] + \frac{1}{2} x'_0 K_i x_0 + q_i(0)' \cdot x_0 + \nu_i(0) \\ & - \frac{1}{2} x(T)' K_{j_T} x(T) - q_{j_T}(T)' \cdot x(T) - \nu_{j_T}(T). \end{aligned}$$

From this we obtain, for every admissible control,

$$(5.4) \quad \begin{aligned} E_{i,x_0} c_T(u) = & \left[E_{i,x_0} \int_0^T \mu_{j_t} dt + E_{i,x_0} \int_0^T \text{tr } G'_{j_t} K_{j_t} G_{j_t} dt \right] \\ & + E_{i,x_0} \int_0^T \|u(t) + R_{j_t}^{-1}B'_{j_t}[K_{j_t}x(t) + q_{j_t}]\|_{R_{j_t}}^2 dt \\ & - E_{i,x_0} [\frac{1}{2}x(T)'K_{j_T}x(T) + q_{j_T}(T)' \cdot x(T) + \nu_{j_T}(T)] \\ & + [\frac{1}{2}x'_0K_ix_0 + q_i(0)' \cdot x_0 + \nu_i(0)], \end{aligned}$$

while, for the feedback control,

$$(5.5) \quad u^*(t) = -R_{j_t}^{-1}B'_{j_t}[K_{j_t}x^*(t) + q_{j_t}]$$

with the response $x^*(\cdot)$, we obtain

$$(5.6) \quad \begin{aligned} E_{i,x_0} c_T(u^*) = & \left[E_{i,x_0} \int_0^T \mu_{j_t} dt + E_{i,x_0} \int_0^T \text{tr } G'_{j_t} K_{j_t} G_{j_t} dt \right] \\ & - E_{i,x_0} \left[\frac{1}{2} x^*(T)' K_{j_T} x^*(T) + q_{j_T}(T)' \cdot x^*(T) + \nu_{j_T}(T) \right] \\ & + \left[\frac{1}{2} x'_0 K_i x_0 + q_i(\nu)' \cdot x_0 + \nu_i(0) \right]. \end{aligned}$$

However, we do not yet know that $u^*(\cdot)$ is an admissible control.

Remark 5.1. The first term in (5.4) will prove to measure the minimal linear growth of the cost. Thus, the minimal cost growth rate is the algebraic sum of two terms. The first is caused by the Markov disturbances and the fact that the trajectory $\Gamma(\cdot)$ cannot be tracked precisely even by a deterministic plant. The second is $\lim_{T \rightarrow \infty} (1/T) E \int_0^T \text{tr } G'_{j_t} K_{j_t} G_{j_t} dt$, which is a consequence of the diffusion noises in the system.

The following result will be invoked in the next section and will also be needed to establish Proposition 5.3 below. Since system (2.5) is stabilizable, it follows by Proposition 2.4 that there exist positive definite matrices $\{K_i\}_{i=1}^N$ satisfying (2.9), and we denote

$$(5.7) \quad F_i = A_i - B_i R_i^{-1} B'_i K_i, \quad i = 1, \dots, N.$$

It follows that the random evolution system

$$(5.8) \quad \dot{x}(t) = F_{j_t} x(t), \quad x(0) = x_0, \quad j(0) = i$$

is stable; that is, $E_{i,x_0} |x(t)|^2 \rightarrow 0$ as $t \rightarrow \infty$ for every x_0 and i . Let \mathbf{Y}_t be the σ -algebra generated by $\{x_s, j_s, 0 \leq s \leq t\}$.

LEMMA 5.2. *Let $t \rightarrow \rho(t)$ be a stochastic process that is adapted to $\{Y_t\}_{t \geq 0}$ and such that $t \rightarrow E_{i,x_0} |\rho(t)|^2$ is a bounded function on $[0, \infty)$. Consider*

$$(5.9) \quad \dot{y}(t) = F_{j_t} y(t) + \rho(t), \quad y(0) = x_0, \quad j_0 = i.$$

Then $t \rightarrow E_{i,x_0} |y(t)|^2$ is a bounded function on $[0, \infty)$.

Proof. It follows from (2.9) that

$$(5.10) \quad F'_i K_i + K_i F_i + \sum_{j=1}^N g_{ij} F_j + M_i = 0, \quad 1 \leq i \leq N$$

for some positive definite matrices M_i , $1 \leq i \leq N$. (In fact, $M_i = Q_i + K_i B_i R_i^{-1} B'_i K_i$.) If $y(\cdot)$ is a solution to (5.9), then

$$\begin{aligned} E_{i,x_0} y(t)' K_{j_t} y(t) = & y'_0 K_i y_0 \\ & - \int_0^T E_{i,x_0} y_{j_s}(s)' M_{j_s} y_{j_s}(s) ds + 2 \int_0^T E_{i,x_0} \rho(s)' K_{j_s} y(s) ds. \end{aligned}$$

We thus get

$$(5.11) \quad \frac{d}{dt} E_{i,x_0} y(t)' K_{j_t} y(t) = -E_{i,x_0} y_{j_t}(t)' M_{j_t} y_{j_t}(t) + 2 E_{i,x_0} \rho(t)' K_{j_t} y(t).$$

We denote $\psi_i(t) = E_{i,x_0} y(t)' K_{j_t} y(t)$. Then the positive definiteness of the matrices M_j , $1 \leq j \leq N$, implies that

$$E_{i,x_0} y(t)' M_{j_t} y(t) \geq \alpha \psi_i(t)$$

for some $\alpha > 0$ and all $1 \leq i \leq N$ and $t \geq 0$. Moreover, since $t \rightarrow E_{i,x_0} |\rho(t)|^2$ is bounded on $[0, \infty)$, there is a $\beta_1 > 0$ such that

$$E_{i,x_0} \rho(t)' K_{j_i} y(t) \leq \beta_1 \left[\sum_{j=1}^N \psi_j(t) \right]^{1/2}$$

for all $t \geq 0$. It thus follows from (5.11) that for some $\beta > 0$

$$\frac{d}{dt} \left[\sum_{i=1}^N \psi_i(t) \right] \leq -\alpha \sum_{i=1}^N \psi_i(t) + \beta \left[\sum_{i=1}^N \psi_i(t) \right]^{1/2},$$

which implies that $t \rightarrow \psi_i(t)$ is a bounded function on $[0, \infty)$ for every $1 \leq i \leq N$. \square

PROPOSITION 5.3. *The feedback control $u^*(\cdot)$ in (5.5) is an admissible control.*

Proof. Since clearly $u^*(\cdot)$ is adapted to $\{\mathbf{Y}_t\}_{t \geq 0}$ all we must show is that the function $t \rightarrow E|x^*(t)|^2$ is bounded on $[0, \infty)$. The process $x^*(\cdot)$ is a solution of the equation

$$(5.12) \quad dx_t^* = [F_{j_i} x^*(t) + \rho(t)] dt + G_{j_i} d\beta_t$$

and that $\rho(\cdot)$ is a bounded process (since each $q_i(\cdot)$ is periodic and bounded). The solution $x^*(\cdot)$ of (5.12) may be written in the form $x^*(t) = y(t) + z(t)$, where $y(\cdot)$ is a solution of (5.9), and $z(\cdot)$ is a solution of

$$(5.13) \quad dz(t) = F_{j_i} z(t) dt + G_{j_i} d\beta_t, \quad z(0) = 0.$$

It follows from Lemma 5.2 that $t \rightarrow E|y(t)|^2$ is bounded on $[0, \infty)$; hence it is enough to prove that $t \rightarrow E|z(t)|^2$ is bounded on $[0, \infty)$.

Let the random fundamental solution of (5.8) be denoted by $\Phi_i(t, s)$, and it follows from the stability of (5.8) that for some $C, \alpha > 0$,

$$(5.14) \quad E|\Phi_i(t, s)|^2 \leq C e^{-\alpha(t-s)}$$

for all $1 \leq i \leq N$ and $0 \leq s \leq t < \infty$. The solution $z(\cdot)$ of (5.13) is $z(t) = \int_0^t \Phi_{j_i}(t, s) G_{j_i} d\beta_s$; hence

$$(5.15) \quad E_i |z(t)|^2 = E_i \operatorname{tr} \int_0^t \phi_{j_i}(t, s) G_{j_i}' G_{j_i} \phi_{j_i}'(t, s) ds.$$

It follows from (5.14) and (5.15) that

$$E_i |z(t)|^2 \leq \gamma \int_0^t e^{-\alpha(t-s)} ds$$

for some constant $\gamma > 0$, implying that $t \rightarrow E_i |z(t)|^2$ is bounded on $[0, \infty)$, which completes the proof of the proposition. \square

THEOREM 5.4. *The control $u^*(\cdot)$ in (5.5) is the unique overtaking optimal control in the class of all the admissible controls, that is,*

$$(5.16) \quad \limsup_{T \rightarrow \infty} [E_i c_T(u^*) - E_i c_T(u)] \leq 0$$

for every admissible control $u(\cdot)$ and every $1 \leq i \leq N$.

Proof. Let $u(\cdot)$ be an admissible control and denote

$$(5.17) \quad v(t) = u(t) + R_{j_i}^{-1} B_{j_i}' [K_{j_i} x(t) + q_{j_i}].$$

We compare the expected cost flows $\{E_{i,x_0} c_T(u)\}_{T \geq 0}$ and $\{E_{i,x_0} c_T(u^*)\}_{T \geq 0}$, which are expressed in (5.4) and (5.6), respectively. Since both $u(\cdot)$ and $u^*(\cdot)$ are admissible, it follows that the function

$$t \rightarrow E_{i,x_0} |x(t)|^2 + E_{i,x_0} |x^*(t)|^2$$

is bounded on $[0, \infty)$. Therefore the difference

$$(5.18) \quad \begin{aligned} E_{i,x_0} c_T(u^*) - E_{i,x_0} c_T(u) &= E_{i,x_0} [\tfrac{1}{2} x(T)' K_T x(T) + q_{j_T}(T)' \cdot x(T) \\ &\quad - \tfrac{1}{2} x^*(T)' K_T x^*(T) - q_{j_T}(T)' \cdot x^*(T)] \\ &\quad - E_{i,x_0} \int_0^T |v(t)|^2 dt \end{aligned}$$

will tend to $-\infty$ if $E_{i,x_0} \int_0^\infty |v(t)|^2 dt = \infty$. Thus it is enough to prove (5.16) for processes $v(\cdot)$ that satisfy

$$(5.19) \quad E_{i,x_0} \int_0^\infty |v(t)|^2 dt < \infty.$$

If $x(\cdot)$ is the response to $u(\cdot)$, then, in view of (5.17), it satisfies the equation

$$(5.20) \quad dx(t) = [F_{j_t} x(t) + \rho(t) + B_{j_t} v(t)] dt + G_{j_t} d\beta(t).$$

We denote $y(t) = x(t) - x^*(t)$ and it follows from (5.12) and (5.20) that $y(\cdot)$ is almost surely absolutely continuous and satisfies the equation

$$\frac{dy}{dt} = F_{j_t} y(t) + B_{j_t} v(t).$$

It follows from Lemma 3.2 that $E_i |y(t)|^2 \rightarrow 0$ as $t \rightarrow 0$, so that

$$\lim_{T \rightarrow \infty} E_i [x(t)' K_{j_T} x(T) - x^*(T)' K_{j_T} x^*(T)] = 0$$

and

$$\lim_{T \rightarrow \infty} E_i q_{j_T}(T)' \cdot [x(T) - x^*(T)] = 0.$$

It follows from (5.17) that $\lim_{T \rightarrow \infty} E_{i,x_0} [c_T(u^*) - c_T(u)]$ exists and is nonpositive, proving (5.16). In fact, we proved that this limit is negative except for the case where $v(\cdot)$ satisfies $E_{i,x_0} \int_0^\infty |v(t)|^2 dt = 0$. Thus the uniqueness of the overtaking optimal control $u^*(\cdot)$ is established, concluding the proof of the theorem. \square

6. Tracking with piecewise diffusion plant and incomplete information. In this section we study the general case of a linear plant that is exposed to modal jumps, nonhomogeneous disturbances, and diffusion noises, and also where the measurement of the state is corrupted by some diffusion noises. Thus the plant is (2.1) with the cost flow process (2.3), but, rather than observing the process $x(\cdot)$, the controller observes the process $y(\cdot)$ in R^k , which is related to $x(\cdot)$ by

$$(6.1) \quad dy(t) = D_{j_t} x(t) dt + H_{j_t} d\tilde{\beta}(t).$$

In (6.1) the process $\tilde{\beta}(\cdot)$ is a \tilde{p} -dimensional Brownian motion, which is assumed to be independent of $\{\beta(t)\}_{t \geq 0}$ and $\{j_t\}_{t \geq 0}$. The matrices $D_i, H_i, 1 \leq i \leq N$, are of appropriate dimensions. Let \mathbf{Y}_t be the σ -algebra generated by $\{j_s, y(s), 0 \leq s \leq t\}$. When the control $u(t) \equiv 0$ is chosen in (2.1), the response $x_0(\cdot)$ and the observed process $y_0(\cdot)$ are obtained, and the corresponding σ -algebra generated by $\{j_s, y_0(s), 0 \leq s \leq t\}$ is denoted by \mathbf{Y}_t^0 . We assume the following.

Assumption 6.1. The piecewise deterministic system

$$(6.2) \quad \dot{x}(t) = A_{j_t} x(t), \quad x(0) = x_0$$

is stable, so that $E|x(t)|^2 \rightarrow 0$ exponentially as $t \rightarrow \infty$, for every $x_0 \in R^n$.

DEFINITION 6.2. The admissible controls are the stochastic process $u(\cdot)$ in R^m such that

- (i) $t \rightarrow u(t)$ is $\{\mathbf{Y}_t\}_{t \geq 0}$ adapted;
- (ii) The function $t \rightarrow E|x(t)|^2$ is bounded on $[0, \infty)$;
- (iii) The σ -algebra \mathbf{Y}_t , which generally contains \mathbf{Y}_t^0 , is, in fact, equal to \mathbf{Y}_t^0 for every $t \geq 0$.

Remark 6.3. Arguing as in the proof of Proposition 5.3, it follows from the stability of (6.2) that the control $u(t) \equiv 0$ is indeed an admissible control. Concerning assumption (iii), the reader is referred to Balakrishnan [3, Chap. 7]. In particular, it is proved there (Theorem 7.1, p. 69) for the case without modal jump disturbances, that controls that are linear transformations of the observed process are such that $\mathbf{Y}_t^0 = \mathbf{Y}_t$. Also, see Willems [26] for a discussion concerning the admissible controls, which are admissible in our framework. Our set of admissible controls contains the locally Lipschitz continuous transformations used, e.g., by Wonham [27] for the linear-quadratic problem.

The assertion in (iii) that $\mathbf{Y}_t^0 \subseteq \mathbf{Y}_t$ is proved in the following result.

LEMMA 6.4. Let $u(\cdot)$ be an admissible control and let $x(\cdot)$ and $y(\cdot)$ be the corresponding response and observed processes in (2.1) and (6.1). Then

- (i) $\mathbf{Y}_t^0 \subseteq \mathbf{Y}_t$ for every $t \geq 0$;
- (ii) Denoting as usual $\hat{x}(t) = E(x(t)|\mathbf{Y}_t)$ then the process $\{x(t) - \hat{x}(t)\}_{t \geq 0}$ does not depend on the choice of the admissible control $u(\cdot)$.

Proof. Let $x_0(\cdot)$ and $y_0(\cdot)$ be the response and the observed process corresponding to the admissible control $u_0(t) \equiv 0$. Let the fundamental solution of (6.2) be $\psi_i(t, s)$. Then, for an admissible control $u(\cdot)$, we denote

$$x_u(t) = \int_0^t \psi_{j_s}(t, s) B_{j_s} u(s) ds.$$

It follows that the response $x(\cdot)$ to $u(\cdot)$ is given by $x(t) = x_0(t) + x_u(t)$, and therefore

$$(6.3) \quad y(t) = y_0(t) + \int_0^t D_{j_s} x_u(s) ds.$$

Since $x_u(s)$ is \mathbf{Y}_s -measurable, it follows from (6.3) that $y_0(t)$ is \mathbf{Y}_t -measurable, for every $t \geq 0$, proving (i) of the lemma.

To prove (ii), we observe that since $u(\cdot)$ is admissible, we indeed have $\mathbf{Y}_t = \mathbf{Y}_t^0$; hence $\hat{x}(t) = \hat{x}_0(t) + x_u(t)$ (which follows from $x = x_0 + x_u$ and the fact that $x_u(t)$ is \mathbf{Y}_t^0 -measurable). This, together with $x(t) = x_0(t) + x_u(t)$, implies that $x(t) - \hat{x}(t) = x_0(t) - \hat{x}_0(t)$, establishing (ii) and concluding the proof of the lemma. \square

The following feedback control

$$(6.4) \quad u_t^* = -R_{j_t}^{-1} B'_{j_t} [K_{j_t} \hat{x}^*(t) + q_{j_t}]$$

will be proved to be the unique overtaking optimal control. This is a straightforward generalization of the optimality of control (5.5) in the complete information case. The optimality of control (6.4) is according to the separation and the certainty equivalence principles.

Let $u(\cdot)$ be an admissible control and now, rather than define $v(\cdot)$ as in (5.16), we define

$$(6.5) \quad w(t) = u(t) + R_{j_t}^{-1} B'_{j_t} [K_{j_t} \hat{x}(t) + q_{j_t}].$$

Then $\{w(t)\}_{t \geq 0}$ is $\{\mathbf{Y}_t\}_{t \geq 0}$ adapted. It follows from (5.3) that

$$\begin{aligned}
 E_i c_T(u) = & E_i \int_0^T \mu_{j_t} dt + \frac{1}{2} E_i \int_0^T \|w(t)\|_{R_{j_t}}^2 dt \\
 & + E_i \int_0^T \text{tr } G'_{j_t} K_{j_t} G_{j_t} dt + \frac{1}{2} x_0 K_i x_0 \\
 (6.6) \quad & + q_i(0)' \cdot x_0 + v_i(0) - E_i [\frac{1}{2} x(T)' K_{j_T} x(T) + q_{j_T}(T)' \cdot x(T) + v_{j_T}(T)] \\
 & + E_{i, x_0} \int_0^T \|B'_{j_t} K_{j_t} [x(t) - \hat{x}(t)]\|_{R_{j_t}^{-1}}^2 dt.
 \end{aligned}$$

To obtain (6.6) we use the fact that $[x(t) - \hat{x}(t)]$ is independent of \mathbf{Y}_t , so that

$$E_i \{ [u(t) + R_{j_t}^{-1} B'_{j_t} (K_{j_t} \hat{x}(t) + q_{j_t})]' \cdot B'_{j_t} K_{j_t} [x(t) - \hat{x}(t)] \} = 0.$$

In particular, for $u^*(\cdot)$ defined in (6.4), expression (6.6) reduces to

$$\begin{aligned}
 E_i c_T(u^*) = & E_i \int_0^T \mu_{j_t} dt + E_i \int_0^T \text{tr } G'_{j_t} K_{j_t} G_{j_t} dt + \frac{1}{2} x_0 K_i x_0 + q_i(0)' \cdot x_0 + v_i(0) \\
 (6.7) \quad & - E_i [x^*(T)' K_{j_T} x^*(T) + q_{j_T}(T)' \cdot x^*(T) + v_{j_T}(T)] \\
 & + E_i \int_0^T \|B'_{j_t} K_{j_t} [x^*(t) - \hat{x}^*(t)]\|_{R_{j_t}^{-1}}^2 dt.
 \end{aligned}$$

By Lemma 6.7 the last terms in the right-hand sides of (6.6) and (6.7) coincide, and it follows in an analogy to (5.17) that

$$\begin{aligned}
 E_{i, x_0} [c_T(u^*) - c_T(u)] = & E_{i, x_0} \frac{1}{2} [x'(T) K_{j_T} x(T) - x^*(T)' K_{j_T} x^*(T)] \\
 (6.8) \quad & + E_{i, x_0} q_{j_T}(T)' \cdot [x(T) - x^*(T)] \\
 & - E_{i, x_0} \int_0^T \|w(t)\|_{R_{j_t}}^2 dt.
 \end{aligned}$$

Our main result is the following theorem.

THEOREM 6.5. *The feedback control $u^*(\cdot)$ defined in (6.4) is the unique overtaking optimal control for system (2.1), (6.1) with cost flow (2.3).*

Proof. The responses $x^*(\cdot)$ to $u^*(\cdot)$ and $x(\cdot)$ to $u(\cdot)$ satisfy the following equations:

$$\begin{aligned}
 dx_t^* = & [F_{j_t} x^*(t) - B_{j_t} R_{j_t}^{-1} B'_{j_t} q_{j_t}] dt \\
 (6.9) \quad & + B_{j_t} R_{j_t}^{-1} B'_{j_t} K_{j_t} [x^*(t) - \hat{x}^*(t)] dt + G_{j_t} d\beta(t),
 \end{aligned}$$

$$\begin{aligned}
 dx_t = & [F_{j_t} x(t) - B_{j_t} R_{j_t}^{-1} B'_{j_t} q_{j_t}] dt + B_{j_t} w(t) dt \\
 (6.10) \quad & + B_{j_t} R_{j_t}^{-1} B'_{j_t} K_{j_t} [x(t) - \hat{x}(t)] dt + G_{j_t} d\beta(t).
 \end{aligned}$$

The matrices F_i , $1 \leq i \leq N$, are such that the piecewise deterministic system (5.8) is stable. Therefore the admissibility of the control $u^*(\cdot)$ will follow from Lemma 5.2 once we have shown that $t \rightarrow E|x^*(t) - \hat{x}^*(t)|^2$ is a bounded function on $[0, \infty)$, by using the same argument as in the proof of Proposition 5.3. By Lemma 6.4 this is equivalent to showing that

$$(6.11) \quad t \rightarrow E|x_0(t) - \hat{x}_0(t)|^2 \text{ is a bounded function on } [0, \infty).$$

Since $x_0(t) - \hat{x}_0(t)$ is independent of \mathbf{Y}_t^0 , while $\hat{x}_0(t)$ is \mathbf{Y}_t^0 -measurable, it follows that

$$E|x_0(t)|^2 = E|x_0(t) - \hat{x}_0(t)|^2 + E|\hat{x}_0(t)|^2,$$

which implies (6.11) in view of the admissibility of $u_0(t) \equiv 0$ (recall Remark 6.3).

If the admissible control $u(\cdot)$ is such that

$$E_i \int_0^\infty \|w(t)\|_{R_i}^2 dt = \infty,$$

then it follows from (6.8) that $\lim_{T \rightarrow \infty} E_i[c_T(u^*) - c_T(u)] = -\infty$, and $u^*(\cdot)$ overtakes $u(\cdot)$. It is thus enough to consider controls $u(\cdot)$, which satisfy

$$(6.12) \quad E_i \int_0^\infty \|w(t)\|_{R_i}^2 dt < \infty.$$

We consider the process $y(t) = x(t) - x^*(t)$, which by (6.9) and (6.10) is almost surely absolutely continuous and satisfies $dy/dt = F_{j_t}y(t) + B_{j_t}w(t)$. It follows from Lemma 3.2 that $E_i|y(t)|^2 \rightarrow 0$ as $t \rightarrow \infty$, from which we conclude that

$$E_i[x(T)'K_{j_T}x(T) - x^*(T)'K_{j_T}x^*(T)] \rightarrow 0$$

and

$$E_i q_{j_T}(T)' \cdot [x(T) - x^*(T)] \rightarrow 0 \quad \text{as } T \rightarrow \infty.$$

It follows from (6.8) that unless $E_i \int_0^\infty \|w(t)\|^2 dt = 0$, the following holds:

$$\lim_{T \rightarrow \infty} E_i[c_T(u^*) - c_T(u)] < 0,$$

which proves that $u^*(\cdot)$ in (6.4) is the unique overtaking optimal control. \square

REFERENCES

- [1] R. A. AKELLA AND P. R. KUMAR, *Optimal control of production rate in failure prone manufacturing system*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 116-126.
- [2] K. J. ASTROM, *Introduction to Stochastic Control Theory*, Academic Press, New York, 1970.
- [3] A. V. BALAKRISHNAN, *Stochastic Differential Systems I*, Springer-Verlag, Berlin, New York, 1978.
- [4] W. A. BROCK, *On existence of weakly maximal programmes in a multisector economy*, Rev. Econom. Stud., 37 (1970), pp. 275-280.
- [5] W. A. BROCK AND A. HAURIE, *On existence of overtaking optimal trajectories over an infinite time horizon*, Math. Oper. Res., 1 (1976), pp. 337-346.
- [6] M. CARAMANIS AND G. LIBEROPOULOS, *Perturbation analysis for the design of flexible manufacturing system flow control*, Mimeo BU-LMSP-88-03, Boston University, Boston, MA, 1988.
- [7] D. A. CARLSON AND A. HAURIE, *Infinite Horizon Optimal Control*, Lecture Notes Econom. Math. Systems, Springer-Verlag, Berlin, New York, 1987.
- [8] K. L. CHUNG, *A Course in Probability Theory*, Academic Press, New York, 1974.
- [9] M. H. A. DAVIS, *Piecewise deterministic Markov processes: A general class of nondiffusion stochastic models*, J. Roy. Statist. Soc., 46 (1984), pp. 353-388.
- [10] W. 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. 1495-1502.
- [11] D. GALE, *On optimal development in multi-sector economy*, Rev. Econom. Stud., 37 (1967), pp. 1-19.
- [12] J. KIMEMIA AND S. B. GERSHWIN, *An algorithm for computer control of a flexible manufacturing system*, IIE Trans., 5 (1983), pp. 353-362.
- [13] A. LEIZAROWITZ, *Optimal trajectories of infinite horizon deterministic control systems*, Appl. Math. Optim., 19 (1989), pp. 11-32.
- [14] ———, *Infinite horizon stochastic regulation and tracking with the overtaking criterion*, Stochastics, 22 (1987), pp. 117-150.
- [15] ———, *Estimates and exact expressions for Lyapunov exponents of stochastic linear differential equations*, Stochastics, 24 (1988), pp. 335-356.

- [16] O. Z. MAIMON AND S. B. GERSHWIN, *Dynamic scheduling and routing for flexible manufacturing systems that have unreliable machines*, Oper. Res., 6 (1988), pp. 279–292.
- [17] T. MOROZAN, *Stabilization of some stochastic discrete time control systems*, Stochastic Anal. Appl., 1 (1983), pp. 89–116.
- [18] ———, *Optimal stationary control for dynamic systems with Markov perturbations*, Stochastic Anal. Appl., 1 (1983), pp. 299–325.
- [19] G. J. OLSDER AND R. SURI, *Time optimal control of parts-routing in a manufacturing system with failure prone machines*, IEEE Internat. Conf. on Decision and Control, Albuquerque, NM, 1980.
- [20] R. RISHEL, *Control of systems with jump disturbances*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 241–244.
- [21] ———, *Dynamic programming and minimum principles for systems with jump Markov disturbances*, SIAM J. Control Optim., 13 (1975), pp. 338–371.
- [22] A. SHARIFNIA, *Production control of a manufacturing system with multiple machine states*, IEEE Trans. Automat. Control, AC-33 (1988), pp. 620–625.
- [23] D. D. SWORDER, *Feedback control of class of linear systems with jump parameters*, IEEE Trans. Automat. Control, AC-14 (1969), pp. 9–14.
- [24] D. VERMES, *Optimal control of piecewise deterministic Markov processes*, Stochastics, 14 (1985), pp. 165–208.
- [25] C. C. VON WEIZSACKER, *Existence of optimal programs of accumulation for an infinite time horizon*, Rev. Econom. Stud., 32 (1965), pp. 85–104.
- [26] J. C. WILLEMS, *The L.Q.G. problem*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. C. Willems, eds., Reidel Publishing, Boston, MA, 1980, pp. 29–44.
- [27] W. M. WONHAM, *Random differential equations in control theory*, in Probabilistic Methods in Applied Mathematics, A. T. Bharucha-Reid, ed., Academic Press, New York, 1970, pp. 131–217.

ACCELERATION OF STOCHASTIC APPROXIMATION BY AVERAGING*

B. T. POLYAK† AND A. B. JUDITSKY‡

Abstract. A new recursive algorithm of stochastic approximation type with the averaging of trajectories is investigated. Convergence with probability one is proved for a variety of classical optimization and identification problems. It is also demonstrated for these problems that the proposed algorithm achieves the highest possible rate of convergence.

Key words. stochastic approximation, recursive estimation, stochastic optimization, optimal algorithms

AMS(MOS) subject classifications. 62L20, 93E10, 93E12

1. Introduction. The methods of stochastic approximation originate in the works [29], [12] and are currently well studied [5], [21], [14], [16], [40]. These methods are widely applied in problems of adaptation, identification, estimation, and stochastic optimization [36], [37], [1], [8]–[10], [18]. The optimal versions (algorithms having the highest rate of convergence) of these methods have been developed as well [34], [38], [6], [27], [28]. However, the application of these optimal methods requires a large amount of a priori information. For example, the matrix $\nabla^2 \ell(x^*)$ must be known in the problem of stochastic optimization (here x^* is the minimum point of $\ell(x)$).

The new way of developing optimal algorithms that does not require such information is based on the idea of averaging the trajectories. It was proposed independently by Polyak [24] and Ruppert [32]. In the latter work, the linear algorithm for the one-dimensional case was considered, and asymptotic normality of the procedure was proved. Polyak [24] studies multidimensional problems and nonlinear algorithms. He has demonstrated the mean square convergence for these methods. In this paper we consider the same framework as in [24], but we demonstrate the asymptotic normality of the estimates. The use of essentially new techniques in the proofs allows us to substantially weaken the conditions of the theorems. Moreover, we prove the statements on almost sure convergence.

The idea of using averaging to accelerate stochastic approximation algorithms appeared in the 1960s (see [36] and the references therein). Afterward, the result was that the hopes associated with this method could not be realized; see, for instance, [23], where it was proved that usual averaging methods are not optimal for linear problems. Nevertheless, the processes with averaging were proposed and studied in the vast variety of papers [11], [14], [20], [13], [33], [4]. The essential advancement [24], [32] was reached on the basis of the paradoxical idea: a slow algorithm having less than optimal convergence rate must be averaged.

The paper is organized as follows. In § 2 the linear case is discussed (i.e., linear equation and linear algorithm). The formulation of the result and proofs are the most clear for that problem. Then in § 3 the general problem of stochastic approximation is studied. The general result obtained is then applied to the unconstrained stochastic optimization problem and to the problem of estimation of linear regression parameters.

2. Linear problem. We want to find x^* , which solves the following equation:

$$(1) \quad Ax = b.$$

Here $b \in R^N$, $x \in R^N$, and $A \in R^{N \times N}$. The sequence $(y_i)_{i \geq 1}$ is observed, where $y_i = Ax_{i-1} - b + \xi_i$. Here $Ax_{i-1} - b$ is a prediction residual and ξ_i is a random disturbance.

* Received by the editors July 30, 1990; accepted for publication (in revised form) June 24, 1991.

† Institute for Control Sciences, Profsoyuznaya 65, 117806, Moscow, Russia.

‡ Institut de Recherche en Informatique et Systemes Aleatoires (IRISA), 35042 Rennes, France.

To obtain the sequence of estimates $(\bar{x}_t)_{t \geq 1}$ of the solution x^* of (1), the following recursive algorithm will be used:

$$(2) \quad \begin{aligned} x_t &= x_{t-1} - \gamma_t y_t, & y_t &= Ax_{t-1} - b + \xi_t, \\ \bar{x}_t &= \frac{1}{t} \sum_{i=0}^{t-1} x_i. \end{aligned}$$

x_0 is an arbitrary (nonrandom) point in R^N .

Let us suppose that the following assumptions hold.

Assumption 2.1. The matrix $-A$ is Hurwitz, i.e., $\operatorname{Re} \lambda_i(A) > 0$. (Here $\lambda_i(A)$ are the eigenvalues of the matrix A .)

Assumption 2.2. Coefficients $\gamma_t > 0$ satisfy either

$$(3) \quad \gamma_t \equiv \gamma, \quad 0 < \gamma < 2 \left(\min_i \operatorname{Re} \lambda_i(A) \right)^{-1}$$

or

$$(4) \quad \gamma_t \rightarrow 0, \quad \frac{\gamma_t - \gamma_{t+1}}{\gamma_t} = o(\gamma_t).$$

Commentary. Condition (4) for $\gamma_t \rightarrow 0$ is the requirement on γ_t to decrease sufficiently slow. For example, the sequences $\gamma_t = \gamma t^{-\alpha}$ with $0 < \alpha < 1$ satisfy this restriction, but the sequence $\gamma_t = \gamma t^{-1}$ does not.

We assume a probability space with an increasing family of Borel fields $(\Omega, \mathfrak{F}, P)$. Suppose that ξ_t is a random variable, adopted to \mathfrak{F}_t .

Assumption 2.3. ξ_t is martingale-difference process, i.e., $E(\xi_t | \mathfrak{F}_{t-1}) = 0$;

$$\sup_t E(|\xi_t|^2 | \mathfrak{F}_{t-1}) < \infty \quad \text{a.s.}$$

(Here $|\cdot|$ is a Euclidean norm in R^N .)

Assumption 2.4. The following limit exists:

$$\lim_{C \rightarrow \infty} \overline{\lim}_{t \rightarrow \infty} E(|\xi_t|^2 I(|\xi_t| > C) | \mathfrak{F}_{t-1}) \stackrel{P}{=} 0.$$

(Here $I(A)$ is the characteristic function of a set A .)

Assumption 2.5. The following hold:

$$\begin{aligned} (a) \quad & \lim_{t \rightarrow \infty} E(\xi_t \xi_t^T | \mathfrak{F}_{t-1}) \stackrel{P}{=} S > 0; \\ (b) \quad & \lim_{t \rightarrow \infty} E \xi_t \xi_t^T = S > 0. \end{aligned}$$

The notation $S > 0$ means that a matrix S is symmetrical and positive definite.

THEOREM 1. (a) Let Assumptions 2.1–2.4, 2.5(a) be satisfied. Then

$$\sqrt{t}(\bar{x}_t - x^*) \xrightarrow{D} N(0, V);$$

i.e., the distribution of normalized error $\sqrt{t}(\bar{x}_t - x^*)$ is asymptotically normal with zero mean and the covariance matrix

$$(5) \quad V = A^{-1} S (A^{-1})^T.$$

(b) If Assumptions 2.1–2.3, 2.5(b) are satisfied, then

$$\lim_{t \rightarrow \infty} E t(\bar{x}_t - x^*)(\bar{x}_t - x^*)^T = V.$$

(c) Let Assumptions 2.1–2.3 be satisfied and let $(\xi_i)_{i \geq 1}$ be mutually independent and identically distributed. Then

$$\bar{x}_t - x^* \rightarrow 0 \quad \text{a.s.}$$

The proofs of the theorems in this paper are in the Appendix.

Part (b) of the theorem was developed in [24], for the case of independent disturbances. Note that Assumption 2.2 on γ_i is significant. If the sequence $\gamma_i = \gamma t^{-1}$ is chosen for algorithm (2) (as is often done for methods using averaging), then the rate of convergence decreases [23].

It was shown in [26] that in the case of independent noises,

$$E(\hat{x}_t - x^*)(\hat{x}_t - x^*)^T \cong t^{-1}V + o(t^{-1})$$

for all linear recursive estimates \hat{x}_t . This asymptotic rate of convergence is achieved by the algorithm

$$(6) \quad x_t = x_{t-1} - t^{-1}A^{-1}y_t.$$

Method (2) provides the same rate of convergence as the optimal linear algorithm. The advantage of this method is that it does not require any knowledge about A and does not use matrix-valued γ_i . Several versions of algorithm (6) use an estimate of matrix A^{-1} , instead of the true value [22], [31]. The significant advantage of these procedures is that they require only the nonsingularity of A (compare to the rather restrictive Assumption 2.1).

3. Nonlinear problem. For nonlinear problems, consider the classical problem of stochastic approximation [21]. Let $R(x): R^N \rightarrow R^N$ be some unknown function. Observations y_i of the function are available at any point $x_{i-1} \in R^N$ and contain the following random disturbances ξ_i :

$$y_i = R(x_{i-1}) + \xi_i.$$

The problem is finding the solution x^* of the equation $R(x) = 0$ by using the observations y_i under the assumption that a unique solution exists.

To solve the problem, we use the following modification of algorithm (2):

$$(7) \quad \begin{aligned} x_t &= x_{t-1} - \gamma_t y_t, & y_t &= R(x_{t-1}) + \xi_t, \\ \bar{x}_t &= \frac{1}{t} \sum_{i=0}^{t-1} x_i, & x_0 &\in R^N. \end{aligned}$$

The first equation in (7) defines the standard stochastic approximation process. Let the following assumptions be fulfilled.

Assumption 3.1. There exists a function $V(x): R^N \rightarrow R^1$ such that for some $\lambda > 0$, $\alpha > 0$, $\varepsilon > 0$, $L > 0$, and all $x, y \in R^N$, the conditions $V(x) \cong \alpha|x|^2$, $|\nabla V(x) - \nabla V(y)| \leq L|x - y|$, $V(x^*) = 0$, $\nabla V(x - x^*)^T R(x) > 0$ for $x \neq x^*$ hold true. Moreover, $\nabla V(x - x^*)^T R(x) \cong \lambda V(x)$ for all $|x - x^*| \leq \varepsilon$.

Assumption 3.2. There exists a matrix $G \in R^{N \times N}$ and $K_1 < \infty$, $\varepsilon > 0$, $0 < \lambda \leq 1$ such that

$$(8) \quad |R(x) - G(x - x^*)| \leq K_1|x - x^*|^{1+\lambda},$$

for all $|x - x^*| \leq \varepsilon$ and $\operatorname{Re} \lambda_i(G) > 0$, $i = \overline{1, N}$.

Assumption 3.3. $(\xi_t)_{t \geq 1}$ is a martingale-difference process, defined on a probability space $(\Omega, \mathcal{F}, \mathcal{F}_t, P)$, i.e., $E(\xi_t | \mathcal{F}_{t-1}) = 0$ almost surely, and for some K_2

$$E(|\xi_t|^2 | \mathcal{F}_{t-1}) + |R(x_{t-1})|^2 \leq K_2(1 + |x_{t-1}|^2) \quad \text{a.s.}$$

for all $t \geq 1$. The following decomposition takes place:

$$(9) \quad \xi_t = \xi_t(0) + \zeta_t(x_{t-1}),$$

where

$$E(\xi_t(0) | \mathcal{F}_{t-1}) = 0 \quad \text{a.s.},$$

$$E(\xi_t(0)\xi_t^T(0) | \mathcal{F}_{t-1}) \xrightarrow{P} S \quad \text{as } t \rightarrow \infty; \quad S > 0,$$

$$\sup_t E(|\xi_t(0)|^2 I(|\xi_t(0)| > C) | \mathcal{F}_{t-1}) \xrightarrow{P} 0 \quad \text{as } C \rightarrow \infty;$$

and, for all t large enough,

$$E(|\zeta_t(x_{t-1})|^2 | \mathcal{F}_{t-1}) \leq \delta(x_{t-1}) \quad \text{a.s.}$$

with $\delta(x) \rightarrow 0$ as $x \rightarrow 0$.

Assumption 3.4. It holds that $(\gamma_t - \gamma_{t+1})/\gamma_t = o(\gamma_t)$, $\gamma_t > 0$ for all t ;

$$(10) \quad \sum_{t=1}^{\infty} (1 + \lambda)/\gamma_t^2 t^{-1/2} < \infty.$$

Commentary. Assumption 3.4, when compared to Assumption 3.2 of Theorem 1, not only restricts the rate of decrease of the coefficients γ_t from above, but it forces the coefficients to decrease not very slowly. Thus, if $\lambda = 1$ in (8), then the sequence $\gamma_t = \gamma t^{-\alpha}$ satisfies this condition only for $\frac{1}{2} < \alpha < 1$.

THEOREM 2. *If Assumptions 3.1–3.4 are satisfied, then $\bar{x}_t \rightarrow x^*$ almost surely, and*

$$\sqrt{t}(\bar{x}_t - x^*) \xrightarrow{D} N(0, V).$$

Here

$$(11) \quad V = G^{-1}S(G^{-1})^T.$$

A proposition similar to Theorem 2 has been stated in [32] for the one-dimensional case. It is well known (see, for example, [6], [21]) that the stochastic approximation algorithm obtains the maximum rate of convergence if it has the form

$$x_t = x_{t-1} - t^{-1}R'(x^*)^{-1}y_t.$$

For that method, $\sqrt{t}(x_t - x^*) \xrightarrow{D} N(0, V)$; here V is the same as in (11). The algorithm, however, could not be realized in that form (the matrix $R'(x^*)$ is unknown). There are some implementable versions of the optimal algorithm [38], [22], [7], [2], [31], but all of them utilize an estimate of the matrix $R'(x^*)$ and usually require additional observations. Algorithm (7) achieves the same optimal rate of convergence and has smaller computational complexity. We must repeat here the comment that already appears at the end of § 2: several procedures that use the estimate of the matrix $R'(x^*)$ [22], [31] do not require the assumption that $\text{Re } \lambda_i(R'(x^*)) > 0$.

4. Stochastic optimization. Consider the problem of searching for the minimum x^* of the smooth function $\ell(x)$, $x \in R^N$. The values of the gradient $y_t = \nabla \ell(x_{t-1}) + \xi_t$

containing random noise ξ_t are available at an arbitrary point x_{t-1} of R^N . To solve this problem, we use the following algorithm of the form (7):

$$(12) \quad \begin{aligned} x_t &= x_{t-1} - \gamma_t \varphi(y_t), & y_t &= \nabla \ell(x_{t-1}) + \xi_t, \\ \bar{x}_t &= \frac{1}{t} \sum_{i=0}^{t-1} x_i, & x_0 &\in R^N. \end{aligned}$$

Let the following assumptions be fulfilled.

Assumption 4.1. Let $\ell(x)$ be a twice continuously differentiable function and $II \leq \nabla^2 \ell(x) \leq LI$ for all x and some $l > 0$ and $L > 0$; here I is the identity matrix.

Assumption 4.2. $(\xi_t)_{t \geq 1}$ is the sequence of mutually independent and identically distributed random variables $E\xi_1 = 0$.

Assumption 4.3. It holds that $|\varphi(x)| \leq K_1(1 + |x|)$.

Assumption 4.4. The function $\psi(x) = E\varphi(x + \xi_1)$ is defined and has a derivative at zero, $\psi(0) = 0$ and $x^T \psi(x) > 0$ for all $x \neq 0$. Moreover, there exist ε , $K_2 > 0$, $0 < \lambda \leq 1$, such that

$$|\psi'(0)x - \psi(x)| \leq K_2|x|^{1+\lambda}$$

for $|x| < \varepsilon$.

Assumption 4.5. The matrix function $\chi(x) = E\varphi(x + \xi_1)\varphi(x + \xi_1)^T$ is defined and is continuous at zero.

Assumption 4.6. The matrix $-G = -\psi'(0)\nabla^2 \ell(x^*)$ is Hurwitz, i.e., $\operatorname{Re} \lambda_i(G) > 0$, $i = \overline{1, N}$.

Assumption 4.7. It holds that $(\gamma_t - \gamma_{t+1})/\gamma_t = o(\gamma_t)$, $\gamma_t > 0$ for all t ;

$$\sum_{t=1}^{\infty} \gamma_t^{(1+\lambda)/2} t^{-1/2} < \infty.$$

The following theorem is a simple corollary of Theorem 2.

THEOREM 3. *Let Assumptions 4.1–4.6 be fulfilled. Then $\bar{x}_t \rightarrow x^*$ almost surely and $\sqrt{t}(\bar{x}_t - x^*) \xrightarrow{D} N(0, V)$, where $V = G^{-1}\chi(0)(G^{-1})^T$.*

The above conditions concerning the function, noises, and score function φ are close to those of [27]. We can find results on the mean square convergence of algorithm (12) under more restrictive conditions (than those of Theorem 3) in [24].

Suppose that disturbance ξ_1 possesses a continuously differentiable density p_ξ and that there exists a finite Fisher information matrix

$$J(p_\xi) = \int (\nabla p_\xi \nabla^T p_\xi) p_\xi^{-1} dy.$$

Let us choose the function φ according to the density p_ξ

$$(13) \quad \varphi(x) = -J^{-1}(p_\xi) \nabla \ln p_\xi(x).$$

In this case, we obtain

$$V = \nabla^2 \ell(x^*)^{-1} J(p_\xi)^{-1} \nabla^2 \ell(x^*)^{-1}.$$

Let us compare the proposed algorithm to the asymptotically optimal form of the stochastic optimization algorithm [27]

$$(14) \quad x_t = x_{t-1} - t^{-1} B \varphi(y_t),$$

where

$$B = \nabla^2 \ell(x^*)^{-1} \quad \text{and} \quad \varphi(y) = -J^{-1}(p_\xi) \nabla \ln p_\xi(y).$$

The value of the matrix $\nabla^2 \ell(x^*)$ is employed in algorithm (14). There exist some implementable versions of the algorithm [39], where an estimate of the matrix is used instead of the true value. Meanwhile, algorithms (12), (13) achieve the same rate of convergence as the optimal unimplementable algorithm (14). Therefore the algorithm with averaging is optimal in this situation in the same sense as in the other problems discussed. Note that this property of optimality corresponds not only to the class of stochastic approximation recursive algorithms, but to a wider class of methods of searching for a minimum point [19].

5. Estimation of regression parameters. Assume that the random variables $x_t \in R^N$, $y_t \in R^1$ are observed in successive instants $t = 1, 2, \dots$, where

$$(15) \quad y_t = x_t^T \theta + \xi_t.$$

Here $\theta \in R^N$ is an unknown parameter and ξ_t is a random noise. We use the following two-step algorithm to produce the sequence of estimates $(\bar{\theta}_t)_{t \geq 1}$ of the parameter θ :

$$(16) \quad \begin{aligned} \theta_t &= \theta_{t-1} + \gamma_t \varphi(y_t - \theta_{t-1}^T x_t) x_t, \\ \bar{\theta}_t &= \frac{1}{t} \sum_{i=0}^{t-1} \theta_i, \quad \theta_0 \in R^N. \end{aligned}$$

Suppose the following assumptions hold true.

Assumption 5.1. Let $(\xi_t)_{t \geq 1}$ be a sequence of mutually independent and identically distributed random variables $E\xi_1 = 0$, $E\xi_1^2 < \infty$.

Assumption 5.2. Let $(x_t)_{t \geq 1}$ be a sequence of mutually independent and identically distributed random variables $E|x_t|^4 < \infty$, $Ex_t x_t^T = B$, $B > 0$. Sequences $(\xi_t)_{t \geq 1}$ and $(x_t)_{t \geq 1}$ are mutually independent.

Assumption 5.3. There exists K_1 such that $|\varphi(x)| \leq K_1(1 + |x|)$ for all $x \in R^N$.

The functions $\psi(x) = E\varphi(x + \xi_1)$, $\chi(x) = E\varphi^2(x + \xi_1)$ are defined under Assumptions 5.1–5.3. Now we state restrictions on ψ , χ .

Assumption 5.4. It holds that $\psi(0) = 0$, $x\psi(x) > 0$ for all $x \neq 0$, $\psi(x)$ has a derivative at zero, and $\psi'(0) > 0$. Moreover, there exist $K_2 < \infty$ and $0 < \lambda \leq 1$ such that

$$|\psi(x) - \psi'(0)x| \leq K_2|x|^{1+\lambda}.$$

Assumption 5.5. The function $\chi(x)$ is continuous at zero.

Assumption 5.6. It holds that $(\gamma_t - \gamma_{t+1})/\gamma_t = o(\gamma_t)$, $\gamma_t > 0$ for all t ;

$$\sum_{t=1}^{\infty} \gamma_t^{(1+\lambda)/2} t^{-1/2} < \infty.$$

THEOREM 4. Assume that Assumptions 5.1–5.6 hold. Then, for algorithm (16), the following properties hold true: $\bar{\theta}_t \rightarrow \theta$ almost surely and $(\bar{\theta}_t - \theta)\sqrt{t} \xrightarrow{D} N(0, V)$, where

$$V = B^{-1} \frac{\chi(0)}{\psi'^2(0)}.$$

The problem of the mean square convergence of method (16) is discussed in [24]. Note that conditions of Theorem 4 are similar to the conditions of standard results for this problem [27]. If ξ_1 possesses a continuously differentiable density function p_ξ , then the optimal algorithm proposed in the latter paper has the following form:

$$(17) \quad \begin{aligned} \theta_t &= \theta_{t-1} + \Gamma_t \varphi(y_t - \theta_{t-1}^T x_t) x_t, \\ \Gamma &= B^{-1} t^{-1}, \quad \varphi(x) = -J(p_\xi)^{-1} p'_\xi(x) / p_\xi(x). \end{aligned}$$

For method (17),

$$(\theta_t - \theta)\sqrt{t} \xrightarrow{D} N(0, V), \quad V = J(p_\xi)^{-1} B^{-1}.$$

Since the matrix B is unknown, algorithm (17) is unimplementable. Nevertheless, it is possible to use, instead of B , its estimate

$$(18) \quad B_t = \left(t^{-1} \sum_{k=1}^t x_k x_k^T \right)^{-1}.$$

In particular, for linear algorithms (i.e., for Gaussian noises) methods (17), (18) coincide with the recursive MLS algorithm. It follows from Theorem 4 that if we choose

$$\varphi(x) = -J(p_\xi)^{-1} p'_\xi(x) / p_\xi(x)$$

for algorithm (16), then the rate of convergence is equal to $V = J(p_\xi)^{-1} B^{-1}$. So the asymptotical rates of convergence of (16) and (17) coincide.

Appendix. Proofs of the theorems consist of the sequence of propositions followed by their proofs. Everywhere in the following, we use the notation $\Delta_t = x_t - x^*$ for an error of the first equation of the algorithm, and $\bar{\Delta}_t = \bar{x}_t - x^*$ for an estimation error. Nonrandom constants that are unimportant will be denoted by the symbols K and α . All relations between random variables are supposed to be true almost surely (unless declared otherwise).

The two matrix lemmas below will be useful in later developments.

Let $(X_j^t)_{t \geq j}$, $(\bar{X}_j^t)_{t \geq j}$ be the sequences of matrices, $\bar{X}_j^t, X_j^t \in R^{N \times N}$, determined by the following recursive relations:

$$(A1) \quad \begin{aligned} X_j^{t+1} &= X_j^t - \gamma_t A X_j^t, & X_j^j &= I, \\ \bar{X}_j^t &= \gamma_j \sum_{i=j}^{t-1} X_j^i. \end{aligned}$$

and $\phi_j^t = A^{-1} - \bar{X}_j^t$.

LEMMA 1. *Let the following hold:*

- (i) *Assumption 2.2 of Theorem 1 holds;*
- (ii) *$\operatorname{Re} \lambda_i(A) > 0$, $i = 1, \bar{N}$.*

Then there is constant $K < \infty$ such that for all j and $t \geq j$

$$(A2) \quad \|\phi_j^t\| \leq K,$$

$$(A3) \quad \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{j=0}^{t-1} \phi_j^t = 0.$$

Proof of Lemma 1.

Part 1. Proposition of the lemma is true with $\gamma_t \equiv \gamma$.

Proof. We obtain from (A1) that

$$\begin{aligned} X_j^t &= (I - \gamma A)^{t-j} & X_j^j &= I, \\ \bar{X}_j^t &= \gamma(I + (I - \gamma A) + \cdots + (I - \gamma A)^{t-j}) = A^{-1} - (I - \gamma A)^{t-j+1} A^{-1}. \end{aligned}$$

The eigenvalues of the matrix $I - \gamma A$ are $\lambda_i(I - \gamma A) = I - \gamma \lambda_i(A)$ and $|\lambda_i(I - \gamma A)| < 1$. So

$$\lim_{t \rightarrow \infty} (I - \gamma A)^t = 0;$$

hence (A2) holds, and

$$\frac{1}{t} \sum_{j=0}^{t-1} \phi_j^t = \frac{1}{t} \sum_{j=0}^{t-1} (I - \gamma A)^{t-j+1} A^{-1} = \frac{1}{t} \sum_{k=2}^{t+1} (I - \gamma A)^k A^{-1} \rightarrow 0 \quad \text{as } t \rightarrow \infty.$$

Part 2. It holds that $t\gamma_t \rightarrow \infty$.

Proof. Let us define $\alpha_t = 1/t\gamma_t$. Then

$$\begin{aligned} \alpha_{t+1} &= (t+1)^{-1} \gamma_{t+1}^{-1} = \frac{1}{t+1} (\gamma_t^{-1} + o(1)) = \alpha_t \frac{t}{t+1} + o(1) \frac{1}{t+1} \\ &= \alpha_t \left(1 - \frac{1}{t+1} \right) + \frac{o(1)}{t+1}, \end{aligned}$$

where $o(1) \rightarrow 0$ as $t \rightarrow \infty$. Since $\sum_{t=1}^{\infty} 1/(t+1) = \infty$, we obtain that $\alpha_t \rightarrow 0$. \square

Part 3. There are $\alpha > 0$ and $K < \infty$ such that for all j and $t \geq j$

$$\|X_j^t\| \leq K \exp \left(-\alpha \sum_{i=j}^{t-1} \gamma_i \right).$$

Proof. From assumption (ii) of the lemma and from the Lyapunov theorem, we have that there exists the solution $V = V^T > 0$ of the Lyapunov equation $A^T V + VA = I$.

Define $L = \max \lambda_i(V)$, $l = \min \lambda_i(V)$, $U_t = (X_j^t)^T V X_j^t$. Then

$$\begin{aligned} (A4) \quad U_{t+1} &= (X_j^t)^T (I - \gamma_t A)^T V (I - \gamma_t A) X_j^t \\ &= U_t - \gamma_t (X_j^t)^T (A^T V + VA) X_j^t + \gamma_t^2 (X_j^t)^T A^T V A X_j^t. \end{aligned}$$

Note that $(X_j^t)^T X_j^t \geq (1/L)(X_j^t)^T V X_j^t$ and $(X_j^t)^T A^T V A X_j^t \leq c(X_j^t)^T V X_j^t$, where $c = (\|A\|^2 L)/l$. Then, for t sufficiently large and some $\lambda > 0$, we get from (A4) that

$$U_{t+1} \leq U_t \left(1 - \frac{1}{L} \gamma_t + c \gamma_t^2 \right) \leq (1 - \lambda \gamma_t) U_t \leq e^{-\lambda \gamma_t} U_t.$$

Thus $U_t \leq U_j \exp(-\lambda \sum_{i=j}^{t-1} \gamma_i)$. However,

$$\|U_t\| \geq l \|X_j^t\|^2 \quad \text{and} \quad \|U_j\| \leq L \|X_j^j\|^2 = L;$$

so we obtain that

$$\|X_j^t\| \leq \sqrt{\frac{L}{l}} \exp \left(-\frac{\lambda}{2} \sum_{i=j}^{t-1} \gamma_i \right). \quad \square$$

Part 4. Equations (A2) and (A3) hold.

Proof. Summing the first equation of (A1) from j to t , we have that

$$(A5) \quad X_j^t = X_j^j - A \sum_{i=j}^{t-1} \gamma_i X_j^i = I - A \sum_{i=j}^{t-1} \gamma_i X_j^i.$$

Let us consider the sum in the right-hand side of (A5). Summing by parts, we get that

$$\sum_{i=j}^{t-1} \gamma_i X_j^i = \gamma_j \sum_{i=j}^{t-1} X_j^i + \sum_{i=j}^{t-1} (\gamma_i - \gamma_j) X_j^i = \bar{X}_j^t + S_j^t.$$

Let us estimate S_j^t . By using the result of Part 3, we obtain that

$$\begin{aligned} (A6) \quad \|S_j^t\| &\leq \left\| \sum_{i=1}^t \left[\sum_{k=j}^{i-1} (\gamma_{k+1} - \gamma_k) \right] X_j^i \right\| \leq \sum_{i=j}^t \sum_{k=j}^{i-1} \gamma_k o(\gamma_k) \|X_j^i\| \\ &\leq o(\gamma_j) \sum_{i=j}^t m_j^i e^{-\lambda m_j^i} = o(\gamma_j) \sum_{i=j}^t \frac{m_j^i e^{-\lambda m_j^i} (m_j^i - m_j^{i-1})}{\gamma_i}, \end{aligned}$$

where $m_j^i = \sum_{k=j}^i \gamma_k$. From Part 2, it follows that $j\gamma_j \leq Ki\gamma_i$ for i sufficiently large. Since $m_j^i = \sum_{k=j}^i \gamma_k \geq \mu(\ln(i/j))$, we can estimate $1/\gamma_i$ as

$$\frac{1}{\gamma_i} \leq K \frac{i}{j\gamma_j} \leq \frac{K}{\gamma_j} \exp\left(\frac{m_j^i}{\mu}\right)$$

for μ arbitrarily large. Finally, we have from (A6) that

$$\|S_j^t\| \leq \frac{Ko(\gamma_j)}{\gamma_j} \sum_{i=j}^t m_j^i e^{-\lambda m_j^i} (m_j^i - m_j^{i-1}) \equiv \frac{Ko(\gamma_j)}{\gamma_j} \int_0^\infty m e^{-\lambda m} dm \varepsilon_j$$

such that, for all $t \geq j$,

$$(A7) \quad \lim_{j \rightarrow \infty} \varepsilon_j = 0.$$

Recall that since $\bar{X}_j' + S_j' = A^{-1} - A^{-1}X_j'$ (see (A5)), we have, by the definition of ϕ_j^t , that

$$\phi_j^t = S_j^t + A^{-1}X_j^t.$$

From Part 3, however, we have that $\|X_j^t\| \leq K$; thus we obtain (A2) from (A7).

Since $\|X_j^t\| \leq K \exp(-\mu(\ln(t/j))) = K(j/t)^\mu$ for μ arbitrarily large, we get that

$$\frac{1}{t} \sum_{j=j_0}^{t-1} \|X_j^t\| \leq K(\mu+1)^{-1}$$

for j_0 large enough. Note that, for some K ,

$$\frac{1}{t} \sum_{j=0}^{t-1} \|X_j^t\| = \frac{1}{t} \sum_{j=0}^{j_0} \|X_j^t\| + \frac{1}{t} \sum_{j=j_0+1}^{t-1} \|X_j^t\| \leq \frac{1}{t} \sum_{j=0}^{j_0} K + \frac{1}{t} \sum_{j=j_0+1}^{t-1} \|X_j^t\|.$$

For arbitrary $\varepsilon > 0$, we can choose μ and $j_0(\mu)$ such that

$$\frac{1}{t} \sum_{j=j_0+1}^{t-1} \|X_j^t\| \leq K(\mu+1)^{-1} \leq \varepsilon/2.$$

Then, choosing t sufficiently large, we get that $1/t \sum_{j=0}^{j_0} K \leq \varepsilon/2$. Hence $1/t \sum_{j=0}^{t-1} \|X_j^t\| \leq \varepsilon$. Moreover, from (A7), we have that

$$\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{j=0}^{t-1} \|S_j^t\| = 0.$$

Hence, from the inequality above, we obtain (A3).

This completes the proof of Lemma 1. \square

Note that we can get from (2) the following equation for the error $\bar{\Delta}_t$ of the algorithm:

$$(A8) \quad \begin{aligned} \Delta_t &= \Delta_{t-1} - \gamma_t(A\Delta_{t-1} + \xi_t), & \Delta_0 &= x_0 - x^*, \\ \bar{\Delta}_t &= \frac{1}{t} \sum_{i=0}^{t-1} \Delta_i. \end{aligned}$$

The next lemma states a convenient representation for the solution of system (A8).

LEMMA 2. *Let the statements of Lemma 1 be fulfilled. Then*

$$(A9) \quad \sqrt{t} \bar{\Delta}_t = \frac{1}{\sqrt{t}\gamma_0} \alpha_t \Delta_0 + \frac{1}{\sqrt{t}} \sum_{j=1}^{t-1} A^{-1} \xi_j + \frac{1}{\sqrt{t}} \sum_{j=1}^{t-1} w_j^t \xi_j,$$

where $\alpha_t, w_j^t \in \mathbb{R}^{N \times N}$ are such that $\|\alpha_t\| \leq K, \|w_j^t\| \leq K$ for some $K < \infty$, and

$$\frac{1}{t} \sum_{j=1}^{t-1} \|w_j^t\| \rightarrow 0 \quad \text{as } t \rightarrow \infty.$$

Proof of Lemma 2. From the first equation of (A8), we have that

$$\Delta_t = \prod_{j=1}^t (I - \gamma_j A) \Delta_0 + \sum_{j=1}^t \prod_{i=j+1}^t (I - \gamma_i A) \gamma_j \xi_j$$

(Set $\prod_{i=n+1}^n (I - \gamma_i A) = I$). Then we get for the error of the algorithm

$$\begin{aligned} \bar{\Delta}_t &= \frac{1}{t} \sum_{j=0}^{t-1} \prod_{i=1}^j (I - \gamma_i A) \Delta_0 + \frac{1}{t} \sum_{k=1}^{t-1} \sum_{j=1}^k \left[\prod_{i=j+1}^k (I - \gamma_i A) \right] \gamma_j \xi_j \\ &= \frac{1}{t} \sum_{j=0}^{t-1} \prod_{i=1}^j (I - \gamma_i A) \Delta_0 + \frac{1}{t} \sum_{j=1}^{t-1} \left[\sum_{k=j}^{t-1} \prod_{i=j+1}^k (I - \gamma_i A) \right] \gamma_j \xi_j. \end{aligned}$$

Set

$$\alpha_j^t = \gamma_j \sum_{i=j}^{t-1} \prod_{k=j+1}^i (I - \gamma_k A),$$

$\alpha_t = \alpha_0^t$, and $w_j^t = \alpha_j^t - A^{-1}$. Then

$$\bar{\Delta}_t = \frac{1}{t\gamma_0} \alpha_t \Delta_0 + \frac{1}{t} \sum_{j=1}^{t-1} A^{-1} \xi_j + \frac{1}{t} \sum_{j=1}^{t-1} w_j^t \xi_j.$$

Note that from (A1) we obtain that $X_j^t = \prod_{i=j}^t (I - \gamma_i A)$ and $\bar{X}_j^t = \gamma_j \sum_{i=j}^t \prod_{k=j}^i (I - \gamma_k A)$.

Thus, from Lemma 1, we get that

$$\lim_{t \rightarrow \infty} \frac{1}{t} \sum_{j=0}^{t-1} \|w_j^t\| = 0, \quad \|w_j^t\| \leq K, \quad \|\alpha_t\| \leq K. \quad \square$$

Proof of Theorem 1.

Part 1. Proposition (a) of the theorem holds.

Proof. We obtain from (A9) that

$$(A10) \quad \sqrt{t} \bar{\Delta}_t = I^{(1)} + I^{(2)} + I^{(3)},$$

where

$$I^{(1)} = \frac{1}{\sqrt{t}} \alpha_t \Delta_0,$$

$$I^{(2)} = \frac{1}{\sqrt{t}} \sum_{j=1}^{t-1} A^{-1} \xi_j,$$

$$I^{(3)} = \frac{1}{\sqrt{t}} \sum_{j=1}^{t-1} w_j^t \xi_j.$$

Note that, since $\|\alpha_t\| \leq K$, $I^{(1)} \rightarrow 0$ in mean square. By Lemma 2 for $I^{(3)}$, we get that

$$E \left| \frac{1}{\sqrt{t}} \sum_{j=1}^{t-1} w_j^t \xi_j \right|^2 \leq \frac{K}{t} \sum_{j=1}^{t-1} \|w_j^t\|^2 \leq \frac{K}{t} \sum_{j=1}^{t-1} \|w_j^t\| \rightarrow 0 \quad \text{as } t \rightarrow \infty,$$

so $|I^{(3)}| \rightarrow 0$. We must demonstrate that the central limit theorem for martingales can be employed for $I^{(2)}$ (see, for example, Theorem 5.5.11 in [17]). We have, for a sufficiently large constant C , that

$$\begin{aligned} & \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{j=1}^{t-1} E(|A^{-1} \xi_j|^2 I(|A^{-1} \xi_j| > C) | \mathcal{F}_{j-1}) \\ & \leq K^2 \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{j=1}^{t-1} E(|\xi_j|^2 I(|\xi_j| > CK^{-1}) | \mathcal{F}_{j-1}) = \ell(C). \end{aligned}$$

According to Assumption 2.4, $\ell(C) \xrightarrow{P} 0$ as $C \rightarrow \infty$. Thus the Lindeberg condition is fulfilled. By Assumption 2.5(a), we get that

$$\frac{1}{t} \sum_{j=1}^{t-1} A^{-1} E(\xi_j \xi_j^T | \mathcal{F}_{j-1})(A^{-1})^T \xrightarrow{P} V.$$

Thus all the conditions of Theorem 5.5.11 [17] are fulfilled.

Part 2. Proposition (b) of the theorem holds.

Proof. We have from (A10) that

$$tE\bar{\Delta}_t\bar{\Delta}_t^T = EI^{(2)}(I^{(2)})^T + \varepsilon_t.$$

As in the proof of Part 1, we obtain from Lemma 2 that $\varepsilon_t \rightarrow 0$ as $t \rightarrow \infty$. Then

$$\begin{aligned} \lim_{t \rightarrow \infty} tE\bar{\Delta}_t\bar{\Delta}_t^T &= \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{j=1}^{t-1} A^{-1} E\xi_j\xi_j^T(A^{-1})^T \\ &= \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{j=1}^{t-1} A^{-1} S(A^{-1})^T = V. \end{aligned}$$

Part 3. Proposition (c) of the theorem holds.

Proof. To simplify notation, we suppose that $\xi_t \in R^1$ (the proof for N -dimensional case is completely analogous). Let us again use decomposition (A10). We immediately get from Lemma 2 that $I^{(1)}/\sqrt{t} \rightarrow 0$. Next, by the law of large numbers (see, e.g., [35]), we get that $I^{(2)}/\sqrt{t} \rightarrow 0$. Let us evaluate the last term of (A10). Define the random sequence $(\bar{\xi}_t)_{t \geq 1}$ by the following equation:

$$\bar{\xi}_t = \begin{cases} \xi_t, & \text{if } |\xi_t| \leq t^{3/4}, \\ 0, & \text{if } |\xi_t| > t^{3/4}. \end{cases}$$

By the Chebyshev inequality, we get that

$$P(|\xi_t| > t^{3/4}) \leq E|\xi_t|^2 t^{-3/2} \leq Kt^{-3/2}.$$

Then $\sum_{i=1}^{\infty} P(|\xi_i| > i^{3/4}) < \infty$ and $P\{|\xi_t| > t^{3/4} \text{ infinitely often}\} = 0$. Since w_j^t are uniformly bounded, it suffices to demonstrate that

$$\frac{1}{t} \sum_{j=1}^{t-1} w_j^t \bar{\xi}_j = t^{-1} S_t \rightarrow 0.$$

Note that $E\xi_t = 0$. Thus

$$\begin{aligned} |E\bar{\xi}_t| &= E\xi_t I(|\xi_t| > t^{3/4}) \leq (E\xi_t^2)^{1/2} (P(|\xi_t| > t^{3/4}))^{1/2} \\ &\leq Kt^{-3/4}. \end{aligned}$$

Then we have that

$$\begin{aligned} S_t^4 &= \left(\sum_{j=0}^{t-1} w_j^t \bar{\xi}_j \right)^4 = \sum_{j=0}^{t-1} (w_j^t)^4 \bar{\xi}_j^4 + K \sum_{\substack{i,j \\ i < j}}^{t-1} (w_j^t)^2 (w_i^t)^2 \bar{\xi}_i^2 \bar{\xi}_j^2 \\ &\quad + K \sum_{\substack{i \neq j \\ i \neq k \\ j < k}}^{t-1} (w_i^t) w_j^t w_k^t \bar{\xi}_i^2 \bar{\xi}_j \bar{\xi}_k \\ &\quad + K \sum_{i < j < k < l}^{t-1} w_i^t w_j^t w_k^t w_l^t \bar{\xi}_i \bar{\xi}_j \bar{\xi}_k \bar{\xi}_l \\ &\quad + K \sum_{i \neq j}^{t-1} w_i^t (w_j^t)^3 \bar{\xi}_i \bar{\xi}_j^3 = \sum_{i=1}^5 I_t^{(i)}. \end{aligned}$$

Note that

$$EI_t^{(1)} = KE \sum_{j=0}^{t-1} \bar{\xi}_j^4 \leq Kt^{3/2} \sum_{j=0}^{t-1} E\bar{\xi}_j^2 \leq Kt^{5/2}.$$

For $I_t^{(5)}$, we have that

$$\begin{aligned} |EI_t^{(5)}| &\leq \left| E \sum_{i \neq j}^{t-1} K\bar{\xi}_i \bar{\xi}_j^3 \right| \leq 2 \left| \sum_{i>j}^{t-1} KE\bar{\xi}_j^3 E\bar{\xi}_i \right| \\ &\leq K \sum_{i>j}^{t-1} j^{3/4} E\bar{\xi}_j^2 i^{-3/4} \leq K \sum_{i>j}^{t-1} E\bar{\xi}_j^2 \leq Kt^2. \end{aligned}$$

By the same arguments, we get that

$$|EI_t^{(2)}| \leq Kt^2, \quad |EI_t^{(3)}| \leq Kt^{3/2}, \quad |EI_t^{(4)}| \leq Kt.$$

Therefore we obtain that

$$t^{-4}ES_t^4 \leq Kt^{-4}(t^{5/2} + t^2 + t^{3/2} + t) \leq Kt^{-3/2}.$$

By the Chebyshev inequality, we get that

$$\sum_{t=1}^{\infty} P(|t^{-1}S_t| > \delta) \leq \sum_{t=1}^{\infty} (t\delta)^{-4}ES_t^4 \leq K \sum_{t=1}^{\infty} t^{-3/2} < \infty.$$

Hence $t^{-1}S_t \rightarrow 0$. \square

Proof of Theorem 2. Let Δ_t be the error of the first equation of (7). Define the function $\bar{R}(x): \mathbb{R}^N \rightarrow \mathbb{R}^N$ by the equation $\bar{R}(x) = R(x - x^*)$.

Part 1. It holds that $V(\Delta_t) \rightarrow V(\omega)$, where $V(\omega)$ is bounded.

Proof. The increment of the function $V_t = V(\Delta_t)$ on one step of algorithm (7) is given by

$$\begin{aligned} V_t &\leq V_{t-1} - \gamma_t \nabla V_{t-1}^T \bar{R}(\Delta_{t-1}) - \gamma_t \nabla V_{t-1}^T \xi_{t-1}(\nabla_{t-1}) \\ &\quad + \frac{L}{2} \gamma_t^2 |\bar{R}(\Delta_{t-1}) + \xi_t(\Delta_{t-1})|^2 \end{aligned}$$

compare with [25, p. 55]. Taking the expectation, conditioned to \mathfrak{F}_{t-1} , by Assumptions 3.2 and 3.3 for some suitable K , we obtain that

$$\begin{aligned} (A11) \quad E(V_t | \mathfrak{F}_{t-1}) &\leq V_{t-1} - \gamma_t \nabla V_{t-1}^T \bar{R}(\Delta_{t-1}) + \gamma_t^2 K(|\Delta_{t-1}|^2 + 1) \\ &\leq V_{t-1}(1 + \gamma_t^2 K) + \gamma_t^2 K - \gamma_t \nabla V_{t-1}^T \bar{R}(\Delta_{t-1}). \end{aligned}$$

From Assumption 3.6 and from Part 2 of Lemma 1, we have that $\sum_{t=1}^{\infty} \gamma_t = \infty$. It can be simply recognized from Assumption 3.6 that $\sum_{t=1}^{\infty} \gamma_t^2 < \infty$; so we obtain by the Robbins-Siegmund theorem [30], that $V_t \rightarrow V(\omega)$.

Since $V \geq \alpha|\Delta|^2$ for some $\alpha > 0$, we have from Part 1 of the proof that $P(\sup_t |\Delta_t| < \infty) = 1$. Thus, for every $\varepsilon > 0$, there exists some $R < \infty$ such that

$$(A12) \quad P\left(\sup_t |\Delta_t| \leq R\right) \geq 1 - \varepsilon.$$

Define the stopping time $\tau_R = \inf\{t \geq 1: |\Delta_t| > R\}$.

Part 2. It holds that $E|\Delta_t|^2 I(\tau_R > t) \leq K\gamma_t$.

Proof. On $\{\tau_R > t\}$ we have from (A11) that

$$\begin{aligned} (A13) \quad E(V_t I(\tau_R > t) | \mathfrak{F}_{t-1}) &\leq E(V_t I(\tau_R > t-1) | \mathfrak{F}_{t-1}) \\ &\leq [V_{t-1}(1 + \gamma_t^2 K) - \alpha \gamma_t V_{t-1}] I(\tau_R > t-1) + \gamma_t^2 K \end{aligned}$$

for some $K, \alpha > 0$. Taking the expectation, we obtain from (A13) that

$$EV_t I(\tau_R > t) \leq EV_{t-1} I(\tau_R > t-1)(1 - \gamma_t \alpha + K\gamma_t^2) + K\gamma_t^2.$$

Finally, by Lemma 2.1.26 [3], we obtain that

$$(A14) \quad EV_t I(\tau_R > t) \leq K\gamma_t.$$

Note that almost sure convergence of the algorithm follows from Parts 1 and 2.

Let us define the process $\bar{\Delta}_t^1$ by the following equations:

$$\begin{aligned} \Delta_t^1 &= \Delta_{t-1}^1 - \gamma_t G \Delta_{t-1}^1 + \gamma_t \xi_t, & \Delta_1^0 &= \Delta_0, \\ \bar{\Delta}_t^1 &= \frac{1}{t} \sum_{i=0}^{t-1} \Delta_i^1. \end{aligned}$$

Let us demonstrate, that for the process $\bar{\Delta}_t^1$, all the properties to be proved follow from Theorem 1.

Part 3. It holds that

$$\lim_{C \rightarrow \infty} \overline{\lim}_{t \rightarrow \infty} E(|\xi_t|^2 I(|\xi_t| > C) | \mathfrak{F}_{t-1}) \stackrel{P}{=} 0.$$

Proof. By decomposition (9), we have that

$$I(|\xi_t| > C) \leq I\left(|\xi_t(\Delta_{t-1})| > \frac{C}{2}\right) + I\left(|\xi_t(0)| > \frac{C}{2}\right);$$

so

$$\begin{aligned} & E(|\xi_t|^2 I(|\xi_t| > C) | \mathfrak{F}_{t-1}) \\ & \leq 2E\left(|\xi_t(\Delta_{t-1})|^2 I\left(|\xi_t(\Delta_{t-1})| > \frac{C}{2}\right) \middle| \mathfrak{F}_{t-1}\right) \\ & \quad + 2E\left(|\xi_t(0)|^2 I\left(|\xi_t(0)| > \frac{C}{2}\right) \middle| \mathfrak{F}_{t-1}\right) \\ & \leq 2\delta(\Delta_{t-1}) + E\left(|\xi_t(0)|^2 I\left(|\xi_t(0)| > \frac{C}{2}\right) \middle| \mathfrak{F}_{t-1}\right) = I_1 + I_2. \end{aligned}$$

Then $I_2 \rightarrow 0$ as $t \rightarrow \infty$ and $C \rightarrow \infty$ by Assumption 2.3; $I_1 \rightarrow 0$, since Δ_t converges to zero.

Therefore all the conditions of proposition (a) of Theorem 1 hold for the process $\bar{\Delta}_t^1$.

We demonstrate the proximity of the processes $\bar{\Delta}_t^1$ and $\bar{\Delta}_t$. Set $\delta_t = \bar{\Delta}_t^1 - \bar{\Delta}_t$; then for δ_t we obtain the equation (compare with (A9))

$$\begin{aligned} \sqrt{t} \delta_t &= \frac{1}{\sqrt{t} \gamma_0} \alpha_t \Delta_0 + \frac{1}{\sqrt{t}} \sum_{j=1}^{t-1} (G^{-1} + w_j')(\bar{R}(\Delta_j) - G\Delta_j) \\ &= I_t^{(1)} + I_t^{(2)}. \end{aligned}$$

Part 4. It holds that $\delta_t \sqrt{t} \rightarrow 0$ as $t \rightarrow \infty$.

Proof. From Lemma 2 we immediately get that $I_t^{(1)} \rightarrow 0$ as $t \rightarrow \infty$. Next, due to Assumption 2.2 and Lemma 2, we get that

$$\begin{aligned} I_t^{(2)} &\leq \sum_{i=0}^{\infty} \frac{1}{i^{1/2}} |(G^{-1} + w_j')(\bar{R}(\Delta_i) - G\Delta_i)| \\ &\leq K \sum_{i=0}^{\infty} \frac{1}{i^{1/2}} |\bar{R}(\Delta_i) - G\Delta_i| \\ &\leq K \sum_{i=0}^{\infty} \frac{|\Delta_i|^{1+\lambda}}{i^{1/2}}. \end{aligned}$$

Thus we obtain from (A14) and Assumption 2.4 that

$$\sum_{i=0}^{\infty} \frac{E(|\Delta_i|^{1+\lambda} I(\tau_R > t))}{i^{1/2}} \leq \sum_{i=0}^{\infty} \frac{K\gamma_i^{(1+\lambda)/2}}{i^{1/2}} < \infty;$$

so

$$\sum_{i=0}^{\infty} \frac{|\Delta_i|^{1+\lambda} I(\tau_R > t)}{i^{1/2}} < \infty.$$

Since

$$\left\{ \sum_{i=0}^{\infty} \frac{|\Delta_i|^{1+\lambda}}{i^{1/2}} < \infty \right\} \supseteq \left\{ \sup_i |\Delta_i| \leq R \right\} \cap \left\{ \sum_{i=0}^{\infty} \frac{|\Delta_i|^{1+\lambda} I(\tau_R > t)}{i^{1/2}} < \infty \right\},$$

we have, by (A12), that

$$P \left\{ \sum_{i=0}^{\infty} \frac{|\Delta_i|^{1+\lambda}}{i^{1/2}} < \infty \right\} \geq 1 - \varepsilon.$$

By the arbitrary choice of ε in (A12),

$$\sum_{i=0}^{\infty} \frac{|\Delta_i|^{1+\lambda}}{i^{1/2}} < \infty.$$

Hence, by the Kronecker lemma,

$$I_t^{(2)} = \frac{1}{\sqrt{t}} \sum_{j=1}^{t-1} \|G^{-1} + w_j'\| |\bar{R}(\Delta_j) - G\Delta_j| \rightarrow 0.$$

So the processes $\bar{\Delta}_t^1$ and $\bar{\Delta}_t$ are asymptotically equivalent.

This completes the proof of Theorem 2. \square

Proof of Theorem 3. Let us check whether the assumptions of Theorem 2 are fulfilled. For that purpose, we transform the first equation of algorithm (12) in the following way:

$$\begin{aligned} (A15) \quad x_t &= x_{t-1} - \gamma_t \psi(\nabla \ell(x_{t-1})) + \gamma_t (\psi(\nabla \ell(x_{t-1})) - \varphi(\nabla \ell(x_{t-1}) + \xi_t)) \\ &= x_{t-1} - \gamma_t R(x_{t-1}) + \gamma_t \xi_t(x_{t-1} - x^*); \end{aligned}$$

here

$$\begin{aligned} (A16) \quad \xi_t(x_{t-1} - x^*) &= \psi(\nabla \ell(x_{t-1})) - \varphi(\nabla \ell(x_{t-1}) + \xi_t), \\ R(x_{t-1}) &= \psi(\nabla \ell(x_{t-1})). \end{aligned}$$

From Assumption 3.4 we have that $R^T(x) \nabla \ell(x) > 0$ for all $x \neq 0$. Let $\ell(x^*) = 0$ for the sake of simplicity. It follows from Assumptions 3.1 and 3.4 that there exist $\alpha > 0$, $\alpha' > 0$, $\varepsilon > 0$ such that

$$R^T(x) \nabla \ell(x) \geq \alpha |\nabla \ell(x)|^2 \geq \alpha' \ell(x)$$

for all $|x - x^*| \leq \varepsilon$; hence $\ell(x)$ is a Lyapunov function for (A15), and all corresponding conditions of Theorem 2 are fulfilled.

So we obtain by Assumption 3.4 that

$$\begin{aligned}
 |R(x) - G(x - x^*)| &= |\psi(\nabla \ell(x)) - \psi'(0)\nabla^2 \ell(x^*)(x - x^*)| \\
 &\leq |\psi(\nabla \ell(x)) - \psi'(0)\nabla \ell(x)| \\
 &\quad + |\psi'(0)\nabla \ell(x) - \psi'(0)\nabla^2 \ell(x^*)(x - x^*)| \\
 &\leq K|\nabla \ell(x)|^{1+\lambda} + \|\psi'(0)\| |\nabla \ell(x) - \nabla^2 \ell(x^*)(x - x^*)| \\
 &\leq K|x - x^*|^{1+\lambda} + K|x - x^*|^2 \leq K|x - x^*|^{1+\lambda}.
 \end{aligned}$$

Hence Assumption 3.2 of Theorem 2 is fulfilled. Next, again using the notation Δ_t for the error of the first equation of (12), we note that $\xi_t(\Delta_{t-1})$ is a martingale-difference process and that

$$E|\xi_t(\Delta_{t-1})|^2 \leq K(1 + |\Delta_{t-1}|^2).$$

So, as concluded in the proof of the Theorem 2 (see Parts 1 and 2), $\Delta_t \rightarrow 0$ and

$$(A17) \quad E|\Delta_t|^2 I(t \leq \tau_R) \leq K\gamma_t.$$

Then, from (A17) by Assumptions 3.5 and 3.4, we have that

$$\begin{aligned}
 |E(\xi_t(\Delta_{t-1})\xi_t(\Delta_{t-1})^T | \mathfrak{F}_{t-1}) - \chi(0)| \\
 \leq K|\chi(\Delta_{t-1}) - \chi(0)| + K|\Delta_{t-1}|^2 \rightarrow 0.
 \end{aligned}$$

Next, we obtain that

$$\begin{aligned}
 E(|\xi_t(\Delta_{t-1})|^2 I(|\xi_t(\Delta_{t-1})| > C) | \mathfrak{F}_{t-1}) \\
 \leq KE(|\xi_t|^2 I(|\xi_t| > C) | \mathfrak{F}_{t-1}) + K|\Delta_{t-1}|^2.
 \end{aligned}$$

From the definition (A16) by Assumption 3.3, we get that

$$I(|\xi_t(\Delta_{t-1})| > C) \leq I(|\Delta_{t-1}| > KC) + I(|\xi_t| > KC);$$

so

$$\begin{aligned}
 E(|\xi_t(\Delta_{t-1})|^2 I(|\xi_t(\Delta_{t-1})| > C) | \mathfrak{F}_{t-1}) \\
 \leq o(1) + KE(|\xi_t|^2 I(|\xi_t| > KC) | \mathfrak{F}_{t-1}) \rightarrow 0 \quad \text{as } t \rightarrow \infty.
 \end{aligned}$$

(Here $o(1) \rightarrow 0$ as $t \rightarrow \infty$.) This means that Assumption 3.3 of Theorem 2 holds. Therefore all conditions of the proposition of Theorem 2 are fulfilled, and the matrix V is defined by the equation

$$V = G^{-1}\chi(0)G^{-1} = (\psi'(0)\nabla^2 \ell(0))^{-1}\chi(0)(\psi'(0)\nabla^2 \ell(0))^{-1}. \quad \square$$

Proof of Theorem 4. Let $\Delta_t = \theta_t - \theta^*$ be an error of the first equation in (16). Denote by \mathfrak{F}_t the minimum σ -algebra generated by disturbances and inputs until the time t : $\mathfrak{F}_t = \sigma(\xi_1, x_1, \dots, \xi_t, x_t)$. Let $R(\Delta) = E\psi(\Delta^T x_1)x_1$. We obtain the following equation for Δ_t :

$$\begin{aligned}
 \Delta_t &= \Delta_{t-1} - \gamma_t E\psi(\Delta_{t-1}^T x_t)x_t \\
 &\quad + \gamma_t (E\psi(\Delta_{t-1}^T x_t)x_t - \varphi(\Delta_{t-1}^T x_t + \xi_t)x_t) \\
 &= \Delta_{t-1} - \gamma_t R(\Delta_{t-1}) + \gamma_t \varepsilon_t,
 \end{aligned} \tag{A18}$$

where $\varepsilon_t = R(\Delta_{t-1}) - \varphi(\Delta_{t-1}^T x_t + \xi_t)x_t$.

We check the fulfillment of the assumptions of Theorem 2 in that case. Assumption 5.4 implies that $\Delta^T R(\Delta) > 0$ for all $\Delta \neq 0$ and $R(\Delta) = 0$ for $\Delta = 0$; so $V(\Delta) = |\Delta|^2$ is the

Lyapunov function for (A18). Hence Assumption 3.1 of Theorem 2 is satisfied. Next, for some K ,

$$|R(\Delta) - \psi'(0)B\Delta| \leq KE|\Delta^T x_1|^{1+\lambda} \leq K(\Delta^T E x_1 x_1^T \Delta)^{(1+\lambda)/2} \leq K|\Delta|^{1+\lambda},$$

and, again, Assumption 3.2 of Theorem 2 holds. Since ε_t is a martingale-difference process and

$$E(|\varepsilon_t|^2 | \mathcal{F}_{t-1}) \leq K(|\Delta_{t-1}|^2 + 1)$$

for some K , we obtain that (see Parts 1 and 2 of the proof of Theorem 2) $|\Delta_t| \rightarrow 0$ and

$$E|\Delta_t|^2 I(t \leq \tau_R) \leq K\gamma_t.$$

By Assumption 5.5, we have that

$$|E(\varepsilon_t \varepsilon_t^T | \mathcal{F}_{t-1}) - \chi(0)B| = |E(\chi(\Delta_{t-1}^T x_t) x_t x_t^T | \mathcal{F}_{t-1}) - \chi(0)B| \xrightarrow{P} 0.$$

We must demonstrate that

$$\sup_t E(|\varepsilon_t|^2 I(|\varepsilon_t| > C) | \mathcal{F}_{t-1}) \xrightarrow{P} 0 \quad \text{as } C \rightarrow \infty.$$

It follows from Assumption 5.3 that

$$\begin{aligned} I(|\varepsilon_t| > C) &\leq I\left(|\varphi(\Delta_{t-1}^T x_t + \xi_t)x_t| > \frac{C}{2}\right) + I\left(|R(\Delta_{t-1})| > \frac{C}{2}\right) \\ &\leq I\left(|\Delta_{t-1}||x_t|^2 > K\frac{C}{2}\right) + I\left(|\xi_t||x_t| > K\frac{C}{2}\right) + I\left(|R(\Delta_{t-1})| > \frac{C}{2}\right) \\ &\leq I\left(|\Delta_{t-1}| > K\sqrt{\frac{C}{2}}\right) + 2I\left(|x_t|^2 > K\sqrt{\frac{C}{2}}\right) + I\left(|\xi_t| > \frac{C}{2}\right) \\ &= I_t^{(1)} + I_t^{(2)} + I_t^{(3)}. \end{aligned}$$

So we obtain that

$$\begin{aligned} E(|\varepsilon_t|^2 I(|\varepsilon_t| > C) | \mathcal{F}_{t-1}) &\leq KE((|\Delta_{t-1}|^2 |x_t|^4 + |\xi_t|^2 |x_t|^2)(I_t^{(1)} + I_t^{(2)} + I_t^{(3)}) | \mathcal{F}_{t-1}) \\ &\leq KI_t^{(1)} |\Delta_{t-1}|^2 + KI_t^{(1)} \\ &\quad + KE(|x_t|^2 I_t^{(2)}) + KE(|\xi_t|^2 I_t^{(3)}) = I_1 + I_2 + I_3 + I_4. \end{aligned}$$

Since $\Delta_t \rightarrow 0$, $I_1 \rightarrow 0$ and $I_2 \rightarrow 0$ as $C \rightarrow \infty$ and $t \rightarrow \infty$. From Assumption 5.1 we get that $I_4 \rightarrow 0$. By the Chebyshev inequality, we get that

$$\begin{aligned} I_3 &\leq K(E|x_t|^4)^{1/2} P^{1/2}(|x_t|^2 > \sqrt{C}) \\ &\leq K \frac{(E|x_t|^4)^{1/2}}{\sqrt{C}} \rightarrow 0 \quad \text{as } C \rightarrow \infty. \end{aligned}$$

So Assumption 3.3 of Theorem 2 is fulfilled. Therefore all the conditions of Theorem 2 are fulfilled under the assumptions of Theorem 4. Finally, we obtain for the matrix V that

$$V = (\psi'(0)B)^{-1} \chi(0)B(\psi'(0)B)^{-1}. \quad \square$$

REFERENCES

- [1] M. A. AIZERMAN, E. M. BRAVERMAN, AND L. I. ROZONER, *The Method of Potential Functions in the Machine Learning Theory*, Nauka, Moscow, 1970. (In Russian.)
- [2] A. M. BENDERSKIY AND M. B. NEVEL'SON, *Multidimensional asymptotically optimal stochastic approximation procedure*, Problems Inform. Transmission, 17 (1982), pp. 423–434.

- [3] A. BENVENISTE, M. METIVIER, AND P. PRIOURET, *Algorithmes Adaptatifs et Approximations Stochastiques (Théorie et Applications)*, Masson, Paris, 1987.
- [4] A. BERMAN, A. FEUER, AND E. WAHNON, *Convergence analysis of smoothed stochastic gradient-type algorithm*, Internat. J. Systems Sci., 18 (1987), pp. 1061–1078.
- [5] YU. M. ERMOL'EV, *Stochastic Programming Methods*, Nauka, Moscow, 1976. (In Russian.)
- [6] V. FABIAN, *Asymptotically efficient stochastic approximation: The RM case*, Ann. Statist., 1 (1973), pp. 486–495.
- [7] ———, *On asymptotically efficient recursive estimation*, Ann. Statist., 6 (1978), pp. 854–866.
- [8] V. N. FOMIN, *Recursive Estimation and Adaptive Filtering*, Nauka, Moscow, 1984. (In Russian.)
- [9] K. S. FU, *Sequential Methods in Pattern Recognition and Machine Learning*, Academic Press, New York, London, 1968.
- [10] G. S. GOODWIN AND K. S. SIN, *Adaptive Filtering, Prediction and Control*, Prentice-Hall, Englewood Cliffs, NJ, 1984.
- [11] A. M. GUPAL AND L. G. BAJENOV, *Stochastic analog of the conjugate gradients method*, Cybernetics, N1 (1972), pp. 125–126. (In Russian.)
- [12] E. KIEFER AND J. WOLFOVITZ, *Stochastic estimation of the maximum of a regression function*, Ann. Math. Statist., 23 (1952), pp. 462–466.
- [13] I. P. KORNFEL'D AND SH. E. SHTEINBERG, *Estimation of the parameters of linear and nonlinear systems using the method of averaged residuals*, Automat. Remote Control, 46 (1986), pp. 966–974.
- [14] A. P. KOROSTELEV, *Stochastic Recurrent Procedures*, Nauka, Moscow, 1981. (In Russian.)
- [15] ———, *On multi-step stochastic optimization procedures*, Automat. Remote Control, 43 (1982), pp. 606–611.
- [16] H. J. KUSHNER AND D. S. CLARK, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer, New York, 1978.
- [17] R. SH. LIPTZER AND A. N. SHIRYAEV, *Martingale Theory*, Nauka, Moscow, 1986. (In Russian.)
- [18] L. LJUNG AND T. SÖDERSTRÖM, *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, MA, 1983.
- [19] A. V. NAZIN, *Informational bounds for gradient stochastic optimization and optimal implemented algorithms*, Automat. Remote Control, 50 (1989), pp. 520–531.
- [20] A. S. NEMIROVSKIY AND D. B. YUDIN, *Complexity of Problems and Effectiveness of Optimization Methods*, Nauka, Moscow, 1980. (In Russian.)
- [21] M. B. NEVEL'SON AND R. Z. KHAS'MINSKIY, *Stochastic Approximation and Recursive Estimation*, American Mathematical Society, Providence, RI, 1973.
- [22] ———, *Adaptive Robbins–Monro procedure*, Automat. Remote Control, 34 (1974), pp. 1594–1607.
- [23] B. T. POLYAK, *Comparison of convergence rate for single-step and multi-step optimization algorithms in the presence of noise*, Engrg. Cybernet., 15 (1977), pp. 6–10.
- [24] ———, *New stochastic approximation type procedures*, Avtomat. i Telemekh., N7 (1990), pp. 98–107. (In Russian.); translated in Automat. Remote Control, 51 (1991), to appear.
- [25] ———, *Introduction to Optimization*, Optimization Software, New York, 1987.
- [26] B. T. POLYAK AND YA. Z. TSYPKIN, *Attainable accuracy of adaptation algorithms*, in Problems of Cybernetics. Adaptive Systems, Nauka, Moscow, 1976, pp. 6–19. (In Russian.)
- [27] ———, *Adaptive estimation algorithms (convergence, optimality, stability)*, Automat. Remote Control, 40 (1980), pp. 378–389.
- [28] ———, *Optimal pseudogradient adaptation algorithms*, Automat. Remote Control, 41 (1981), pp. 1101–1110.
- [29] H. ROBBINS AND S. MONROE, *A stochastic approximation method*, Ann. Math. Statist., 22 (1951), pp. 400–407.
- [30] H. ROBBINS AND D. SIEGMUND, *A convergence theorem for nonnegative almost supermartingales and some applications*, in Optimizing Methods in Statistics, J. S. Rustaji, ed., Academic Press, New York, 1971, pp. 233–257.
- [31] D. RUPPERT, *A Newton–Rafson version of the multivariate Robbins–Monro procedure*, Ann. Statist., 13 (1985), pp. 236–245.
- [32] ———, *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.
- [33] A. RUSZYNSKI AND W. SYSKI, *Stochastic approximation method with gradient averaging for unconstrained problems*, IEEE Trans., AC-28 (1983), pp. 1097–1105.
- [34] D. T. SAKRISON, *Stochastic approximation: A recursive method for solving regression problems*, in Advances in Communication Theory and Applications, 2, A. V. Balakrishnan, ed., Academic Press, New York, London, 1966, pp. 51–106.
- [35] A. N. SHIRYAEV, *Probability*, Nauka, Moscow, 1980. (In Russian.)

- [36] YA. Z. TSYPKIN, *Adaptation and Learning in Automatic Systems*, Academic Press, New York, London, 1971.
- [37] ———, *Foundations of Informational Theory of Identification*, Nauka, Moscow, 1984. (In Russian.)
- [38] J. H. VENTER, *An extension of the Robbins–Monro procedure*, Ann. Math. Statist., 38 (1967), pp. 181–190.
- [39] M. L. VIL'K AND S. V. SHIL'MAN, *Convergence and optimality of implementable, adaptation algorithms (informational approach)*, Problems Inform. Transmission, 20 (1985), pp. 314–326.
- [40] M. WASAN, *Stochastic Approximation*, Cambridge University Press, London, 1970.

STRONGLY AND WEAKLY RELAXED CONTROLS FOR TIME DELAY SYSTEMS*

JAVIER F. ROSENBLUETH†

Abstract. In recent articles, two different relaxation procedures for time delay problems in optimal control have been proposed. A “weak” procedure, due to Warga [*Nonadditively coupled delayed controls*, privately circulated, 1986], for which the existence of minimizers is assured, applies to fully nonlinear problems with delays in the state and control variables. Nevertheless, an example is found for which the effect of weak relaxation reduces the infimum cost. The example involves two delays, one being twice the value of the other, and it belongs to a class of problems that is called “commensurate,” where the quotient of any two delays is rational. For these problems a “strong” procedure is proposed and it is proved that the extended problem has a solution and the extension is “proper”; i.e., the infimum costs coincide. In this paper it is shown through several examples why, in general, the weak relaxation technique may fail to provide a proper extension and, for noncommensurate delay problems, an approximation result in terms of strongly relaxed controls is given.

Key words. time delay systems, relaxation procedures

AMS(MOS) subject classifications. 49A10, 49A50, 49D20

1. Introduction. We are interested in relaxing an optimal control problem to ensure existence of minimizers, but the relaxation technique should also provide a methodology for finding ordinary admissible processes that come close to achieving the infimum cost. This is obtained if the infimum cost of the original problem coincides with the minimum cost of the relaxed version, for then we solve the latter and approximate the relaxed minimizer by an ordinary one. If this is the case, we call the relaxed problem a “proper” extension of the original one. It is well known that, for delay-free problems, the usual relaxation technique provides a proper extension (see [6]).

In a recent paper (see [8]), Warga proposes a relaxation procedure for optimal control problems involving transformations of the state and control functions showing that the resulting relaxed problem has a solution. Properness is not proved and, in [4], we show through an example that this procedure may reduce the infimum cost, thus failing to give a proper extension of the original problem. A new relaxation procedure is proposed in [4] for which the extension is proper in certain situations where Warga’s extension is not. However, for the general case, the question of properness remains unanswered.

To clearly situate the contributions of this paper, let us summarize the main results obtained so far. We address the following problem:

$$\begin{aligned} & \text{minimize } g(x(1)) \\ \text{(P)} \quad & \text{subject to } \dot{x}(t) = f(t, x(t), u(t), u(t - \theta_1), \dots, u(t - \theta_k)) \quad \text{a.e. in } T, \\ & x(0) = \xi, \quad u(t) \in \Omega \quad \text{a.e. in } [-\theta_k, 1], \end{aligned}$$

where u is any measurable function mapping $[-\theta_k, 1]$ to \mathbf{R}^m , $T := [0, 1]$, and we are given real numbers $0 < \theta_1 < \dots < \theta_k \leq 1$, a point $\xi \in \mathbf{R}^n$, a set $\Omega \subset \mathbf{R}^m$, functions g mapping \mathbf{R}^n to \mathbf{R} , and f mapping $T \times \mathbf{R}^n \times \mathbf{R}^{m(k+1)}$ to \mathbf{R}^n .

* Received by the editors August 6, 1990; accepted for publication (in revised form) March 13, 1991.

† Departamento de Métodos Matemáticos y Numéricos, IIMAS-UNAM, Apartado Postal 20-726, Admón. 20, México, D.F., 01000, México.

This system reflects the main difficulties encountered in finding a proper extension. No further problems are caused if delay terms are also present in the state variable (see [7]) or if we are given endpoint constraints (see [5]).

We assume that the functions and sets delimiting the problem satisfy the following hypotheses:

- (i) f and g are continuous and Ω is compact;
- (ii) there exists an integrable function $\phi : T \rightarrow \mathbf{R}$ such that, for all $(t, r) \in T \times \Omega^{k+1}$ and $x, y \in \mathbf{R}^n$,

$$|f(t, x, r) - f(t, y, r)| \leq \phi(t)|x - y|.$$

We use the following notation. For any $S \subset \mathbf{R}$ compact and R compact metric space, let

$$\mathcal{U}(S, R) := \{u : S \rightarrow R \mid u \text{ is measurable}\}$$

and

$$\mathcal{M}(S, R) := \{\mu : S \rightarrow (\text{frm}(R), |\cdot|_w) \mid \mu \text{ is measurable and } \mu(t) \in \text{rpm}(R) \text{ a.e. in } S\},$$

where $\text{frm}(R)$ denotes the vector space of all Radon (finite regular Borel) measures in R , and $\text{rpm}(R)$ denotes the set of Radon probability measures in R . Since $\text{frm}(R)$ and $C(R)^*$, the dual of the space of continuous functions on R to \mathbf{R} , are isomorphic (Riesz theorem) and $C(R)$ is separable, we consider $\text{frm}(R)$ as a normed space with a weak norm (by Bishop's theorem). Also, $\mathcal{M}(S, R)$ is seen as a normed space with a weak norm and we embed $\mathcal{U}(S, R)$ as a subset of $\mathcal{M}(S, R)$ by identifying each $u \in \mathcal{U}(S, R)$ with the function $t \rightarrow \delta_{u(t)}$, where δ_r denotes the Dirac measure at r .

Now, problem (P) is posed over all admissible ordinary processes; that is, pairs (x, u) comprising an ordinary control u (an element of $\mathcal{U}([- \theta_k, 1], \Omega)$) and an absolutely continuous function x , which satisfies the differential equation (with respect to u) together with the initial-point condition. For delay-free problems (regarded as a special case of (P)), an admissible relaxed process is a pair (x, μ) , where μ is a relaxed control (an element of $\mathcal{M}(T, \Omega)$) and x is an absolutely continuous function satisfying the differential equation

$$\begin{aligned} \dot{x}(t) &= f(t, x(t), \mu(t)) \\ &= \int f(t, x(t), r) \mu(t)(dr) \quad \text{a.e. in } T \end{aligned}$$

together with $x(0) = \xi$ (under the hypotheses, given μ , there exists a unique x such that (x, μ) is an admissible relaxed process). For delay-free problems, it is known that the relaxed problem, that is, (P) posed over admissible relaxed processes, has a minimizer and this extension is proper (see [6]).

For problem (P) involving delays, we can easily find an example, as Warga shows in [8], for which $\mathcal{M}([- \theta_k, 1], \Omega)$ does not provide a proper extension to (P), so that it is no longer sufficient to consider the relaxed version of the control function. The example treated in [8] has one delay and, for this example, Warga shows that a proper extension is provided by a subset of $\mathcal{M}(T, \Omega^2)$. To describe this relaxation procedure, suppose that we are given an admissible ordinary process (x, u) . Set $\theta_0 := 0$ and define $\Delta_i := \theta_i - \theta_{i-1}$, $T_i := [\Delta_i, 1]$, and $u_i(t) := u(t - \theta_i)$ for all $t \in T$ and $i = 1, \dots, k$. Then

(x, \tilde{u}) , with $\tilde{u} = (u_0, u_1, \dots, u_k)$, satisfies

$$\dot{x}(t) = f(t, x(t), \tilde{u}(t)) \quad \text{a.e. in } T,$$

$$x(0) = \xi,$$

$$\tilde{u}(t) \in \Omega^{k+1} \quad \text{a.e. in } T,$$

together with the compatibility conditions

$$u_i(t) = u_{i-1}(t - \Delta_i) \quad \text{a.e. in } T_i, \quad i = 1, 2, \dots, k.$$

Warga's idea is to use the standard theory to define relaxed controls for this system, treating u_0, u_1, \dots, u_k as independent functions. A relaxed control is then a measurable function with values regular probability measures on the Borel sets of Ω^{k+1} , and we impose some conditions generalizing these last compatibility conditions.

Let us denote by $\mathcal{U}(\theta_1, \dots, \theta_k)$ the set of ordinary controls for this system, i.e., the set of measurable functions (u_0, \dots, u_k) in $\mathcal{U}(T, \Omega^{k+1})$ satisfying

$$u_i(t) = u_{i-1}(t - \Delta_i) \quad \text{a.e. in } T_i, \quad i = 1, 2, \dots, k.$$

Warga's extension of $\mathcal{U}(\theta_1, \dots, \theta_k)$, which we call "weak" relaxation, consists of the set of relaxed controls μ in $\mathcal{M}(T, \Omega^{k+1})$ satisfying

$$\mathcal{P}_i \mu(t) = \mathcal{P}_{i-1} \mu(t - \Delta_i) \quad \text{a.e. in } T_i, \quad i = 1, 2, \dots, k,$$

where $\mathcal{P}_i : C^*(\Omega^{k+1}) \rightarrow C^*(\Omega)$, and $\mathcal{P}_i \mu(t)$ denotes the projection onto the i th coordinate of $\mu(t)$. We denote this set by $\mathcal{M}_w(\theta_1, \dots, \theta_k)$, which, clearly, generalizes the compatibility conditions on the original controls.

In [8], Warga proves that problem (P) posed over admissible weakly relaxed processes has a minimizer, and in [4] we show that, for one-delay systems, this extension is proper. For systems involving two delays, things change. We show through an example in [4] that, for any $0 < \theta \leq \frac{1}{2}$, $\mathcal{M}_w(\theta, 2\theta)$ may strictly reduce the infimum cost, thus failing to give a proper extension. For the special case when, for all $i = 1, \dots, k$, each θ_i is of the form $i\theta$ for some $\theta \in (0, 1/k]$, a relaxation procedure, reducing the system into one without delays, was first established by Warga in [6]. The basic idea is to section ordinary controls and the corresponding trajectories into segments of length θ and to stack these segments to form higher-dimensional vector-valued functions on the interval $[0, \theta]$. The resulting functions satisfy a delay-free differential equation, together with a set of mixed boundary conditions that express the equality of one component of the enlarged state function at the endpoint with the initial value of the next component to ensure the continuity of the original trajectory.

Although the usual relaxation procedure for the reduced problem does yield a proper extension (see [6], [1], and [4] for details), it may be unsatisfactory in certain respects. As we mention in [4], the dimension of the spaces involved in the reduced problem can be very large depending on the number of delays involved and the number of segments of length θ . Also, the dimension of the spaces increases rapidly with the length of the underlying time interval. Apart from this, in passing to the reduced problem on the time interval $[0, \theta]$, the connections with the original problem are somewhat obscured. A new relaxation procedure, which we call "strong," is introduced in [4], solving the difficulties that the relaxation via the reduced problem presents.

Observe that, for this case, $\mathcal{U}(\theta, 2\theta, \dots, k\theta)$ is given by the set of functions (u_0, \dots, u_k) in $\mathcal{U}(T, \Omega^{k+1})$ satisfying

$$(u_1, \dots, u_k)(t) = (u_0, \dots, u_{k-1})(t - \theta) \quad \text{a.e. in } [\theta, 1].$$

Strongly relaxed controls are defined as those elements μ of $\mathcal{M}(T, \Omega^{k+1})$ satisfying

$$\mathcal{P}_{1, \dots, k} \mu(t) = \mathcal{P}_{0, \dots, k-1} \mu(t - \theta) \quad \text{a.e. in } [\theta, 1],$$

and in [4] we show that problem (P) posed over admissible strongly relaxed processes has the desired properties: there exists a minimizer and the extension is proper.

This is our starting point. In this paper we analyze the set of weakly relaxed controls and show through several examples why, for systems with more than one delay, it may fail, in general, to give a proper extension. We also give an “approximation” result for general delay systems in terms of strongly relaxed controls.

2. Notation and preliminary results. Since we are dealing with systems defined in different spaces, it will be convenient to introduce the following notation. Denote by X the space of all absolutely continuous functions mapping T to \mathbf{R}^n . If R is any compact metric space, h a function mapping $T \times \mathbf{R}^n \times R$ to \mathbf{R}^n , and \mathcal{S} any subset of $\mathcal{M}(T, R)$, let $\mathcal{A}(h, \mathcal{S})$ be the set of admissible processes with respect to h and \mathcal{S} , that is, those pairs (x, u) in $X \times \mathcal{S}$ satisfying the differential equation

$$\dot{x}(t) = h(t, x(t), u(t)) \quad \text{a.e. in } T,$$

together with the initial condition $x(0) = \xi$. For this system, denote the reachable set by $\mathcal{R}(h, \mathcal{S})$, that is, the set of points $x(1)$ in \mathbf{R}^n such that, for some $u \in \mathcal{S}$, $(x, u) \in \mathcal{A}(h, \mathcal{S})$. By $P(\mathcal{A}(h, \mathcal{S}))$, we mean problem (P) posed over $\mathcal{A}(h, \mathcal{S})$, i.e., the problem of minimizing g over $\mathcal{R}(h, \mathcal{S})$.

Our original problem (P) is posed over pairs belonging to $X \times \mathcal{U}([-\theta_k, 1], \Omega)$ and, describing Warga’s relaxation procedure, we transformed this system into one defined on $X \times \mathcal{U}(T, \Omega^{k+1})$. It is a simple fact to show that both systems are equivalent in the sense that the respective reachable sets coincide. Explicitly, if we set

$$\mathcal{D}_k := \{(\theta_1, \dots, \theta_k) \in \mathbf{R}^k \mid 0 < \theta_1 < \dots < \theta_k \leq 1\}$$

for all $k \in \mathbf{N}$, then the following result holds.

LEMMA 2.1. *For any $k \in \mathbf{N}$ and $(\theta_1, \dots, \theta_k) \in \mathcal{D}_k$, the reachable set for problem (P) coincides with $\mathcal{R}(f, \mathcal{U}(\theta_1, \dots, \theta_k))$.*

In view of this lemma, we may regard our original problem as that of minimizing $g(x(1))$ subject to

$$\dot{x}(t) = f(t, x(t), \tilde{u}(t)) \quad \text{a.e. in } T, \quad x(0) = \xi,$$

and $\tilde{u} \in \mathcal{U}(\theta_1, \dots, \theta_k)$. An extension of this problem corresponds to a problem posed over $\mathcal{A}(f, \mathcal{S})$, where \mathcal{S} is a set containing $\mathcal{U}(\theta_1, \dots, \theta_k)$ and, clearly, the extension will be proper if $\mathcal{R}(f, \mathcal{U}(\theta_1, \dots, \theta_k))$ is dense in $\mathcal{R}(f, \mathcal{S})$.

For any $k \in \mathbf{N}$ and $\theta \in (0, 1/k]$, denote by $\mathcal{M}_k(\theta)$ the set of strongly relaxed controls, an extension of $\mathcal{U}(\theta, 2\theta, \dots, k\theta)$, which we now write as $\mathcal{U}_k(\theta)$. It is clear that $\mathcal{M}_k(\theta)$ is contained in $\mathcal{M}_w(\theta, 2\theta, \dots, k\theta)$ and that, for one-delay systems, both sets coincide. Using this notation, the main results obtained in [4] can be summarized as follows.

THEOREM 2.2. *For any $k \in \mathbf{N}$ and $\theta \in (0, 1/k]$, $P(\mathcal{A}(f, \mathcal{M}_k(\theta)))$ has a minimizer and $\mathcal{R}(f, \mathcal{U}_k(\theta))$ is dense in $\mathcal{R}(f, \mathcal{M}_k(\theta))$.*

THEOREM 2.3. *For any $\theta \in (0, 1/2]$, we can find a function f for which the problem $P(\mathcal{A}(f, \mathcal{M}_w(\theta, 2\theta)))$ does not provide a proper extension of $P(\mathcal{A}(f, \mathcal{U}_2(\theta)))$.*

3. The commensurate case. In this section we analyze systems with commensurate time delays, that is, systems for which θ_i/θ_{i+1} is a rational number for all $i = 1, \dots, k-1$. We study problem (P) involving only two delays, since no difficulties arise in extending the following theory to apply to control problems with three or more delays. Thus, $k=2$, and we assume the existence of positive integers p and q such that $q\theta_1 = p\theta_2$. Note that if we set $\theta := \theta_1/p = \theta_2/q$, then (P) is posed over $\mathcal{A}(f, \mathcal{U}(p\theta, q\theta))$. Let us begin by expressing this system in terms of $\mathcal{U}_q(\theta)$.

Given the original function $f: T \times \mathbf{R}^n \times \mathbf{R}^{3m} \rightarrow \mathbf{R}^n$, define, for any $p, q \in \mathbf{N}$ with $p < q$, the function $\tilde{f}_{p,q}$ mapping $T \times \mathbf{R}^n \times \mathbf{R}^{m(q+1)}$ to \mathbf{R}^n by

$$\tilde{f}_{p,q}(t, x, u_0, u_1, \dots, u_q) := f(t, x, u_0, u_p, u_q)$$

for all $(t, x, u_0, u_1, \dots, u_q) \in T \times \mathbf{R}^n \times \mathbf{R}^{m(q+1)}$. Note that given $p, q \in \mathbf{N}$ with $p < q$ and $\theta \in (0, 1/q]$, we clearly have $\mathcal{R}(\tilde{f}_{p,q}, \mathcal{U}_q(\theta)) \subset \mathcal{R}(f, \mathcal{U}(p\theta, q\theta))$. Conversely, if $(x, u_0, u_1, u_2) \in \mathcal{A}(f, \mathcal{U}(p\theta, q\theta))$ and we define $\tilde{u}_0(t) := u_0(t)$ for all $t \in T$ and, for all $i = 1, \dots, q$,

$$\tilde{u}_i(t) := \begin{cases} \tilde{u}_{i-1}(t - \theta), & t \in [\theta, 1], \\ u_2(t + (q-i)\theta), & t \in [0, \theta), \end{cases}$$

then it is easily seen that $(x, \tilde{u}_0, \tilde{u}_1, \dots, \tilde{u}_q) \in \mathcal{A}(\tilde{f}_{p,q}, \mathcal{U}_q(\theta))$. Thus, we have the following result.

LEMMA 3.1. *For any $p, q \in \mathbf{N}$ with $p < q$ and $\theta \in (0, 1/q]$,*

$$\mathcal{R}(f, \mathcal{U}(p\theta, q\theta)) = \mathcal{R}(\tilde{f}_{p,q}, \mathcal{U}_q(\theta)).$$

Combining this lemma with Theorem 2.2, we obtain the following result, which solves the question of properness for commensurate delay problems.

THEOREM 3.2. *For any positive integers $p < q$ and $\theta \in (0, 1/q]$, $\mathcal{R}(f, \mathcal{U}(p\theta, q\theta))$ is dense in $\mathcal{R}(\tilde{f}_{p,q}, \mathcal{M}_q(\theta))$.*

We now pose the question of whether this procedure may be simplified, for certain cases, in terms of weakly relaxed controls. As we have already mentioned, this is not possible if θ_2 is twice the value of θ_1 , that is, when $p = 1$ and $q = 2$. The example given in [4] is based on the fact that we can find weakly relaxed minimizers that do not satisfy the strong compatibility conditions. We then construct a problem for which the minimum cost for weakly relaxed processes is strictly less than the one for strongly relaxed processes, which by Theorem 2.2 coincides with the infimum cost for the original problem.

If θ_2 is not equal to $2\theta_1$, we might suspect that the weakly relaxed problem will provide a proper extension of the original one. The reason for this can be illustrated through an example. Suppose that $p = 2$ and $q = 3$. Weakly relaxed controls for $P(\mathcal{A}(f, \mathcal{U}(2\theta, 3\theta)))$ are measurable functions with values regular probability measures on the Borel sets of Ω^3 , i.e., elements of $\mathcal{M}(T, \Omega^3)$, and they satisfy the conditions

$$\mathcal{P}_1\mu(t) = \mathcal{P}_0\mu(t - 2\theta) \quad \text{a.e. in } [2\theta, 1],$$

$$\mathcal{P}_2\mu(t) = \mathcal{P}_1\mu(t - \theta) \quad \text{a.e. in } [\theta, 1].$$

Transforming this problem into the equivalent $P(\mathcal{A}(\tilde{f}_{2,3}, \mathcal{U}_3(\theta)))$, a strongly relaxed control σ is a member of $\mathcal{M}(T, \Omega^4)$ satisfying

$$\mathcal{P}_{1,2,3}\sigma(t) = \mathcal{P}_{0,1,2}\sigma(t - \theta) \quad \text{a.e. in } [\theta, 1].$$

Clearly, in view of Theorem 3.2, the weakly relaxed problem will be a proper extension of the original if, for any $\mu \in \mathcal{M}_w(2\theta, 3\theta)$, we can find $\sigma \in \mathcal{M}_3(\theta)$ such that

$$\mathcal{P}_{0,2,3}\sigma(t) = \mu(t),$$

for then the strong compatibility conditions for σ coincide with the weak conditions for μ , and the two problems $P(\mathcal{A}(f, \mathcal{M}_w(2\theta, 3\theta)))$ and $P(\mathcal{A}(\tilde{f}_{2,3}, \mathcal{M}_3(\theta)))$ are equivalent.

This situation may very well hold. For example, if $\Omega = [0, 1]$,

$$\mu(t) = \frac{1}{2} \delta_{(1,1,0)} + \frac{1}{2} \delta_{(0,0,1)} \quad (t \in T)$$

and

$$\sigma(t) = \frac{1}{2} \delta_{(1,0,1,0)} + \frac{1}{2} \delta_{(0,1,0,1)} \quad (t \in T),$$

then, for any $\theta \in (0, 1/3]$, we have $\mu \in \mathcal{M}_w(2\theta, 3\theta)$, $\sigma \in \mathcal{M}_3(\theta)$, and $\mathcal{P}_{0,2,3}\sigma(t) = \mu(t)$. Nevertheless, we can easily find an example for which this is no longer true. If

$$\mu(t) = \frac{1}{2} \delta_{(1,0,1)} + \frac{1}{2} \delta_{(0,1,0)} \quad (t \in T),$$

then $\mu \in \mathcal{M}_w(2\theta, 3\theta)$, but we cannot exhibit a strongly relaxed control $\sigma \in \mathcal{M}_3(\theta)$ for which $\mathcal{P}_{0,2,3}\sigma(t) = \mu(t)$.

Based on these ideas, several examples in this section are given of commensurate delay problems for which weakly relaxed controls fail to provide a proper extension.

For all p and q relatively primes with $1 \leq p < q$, consider the problem, which we label $(P_{p,q})$, of minimizing $x_1(1)$ subject to

$$\dot{x}_1(t) = (x_0(t) - t/2)^2 + h(u(t), u(t - p\theta), u(t - q\theta)) \quad \text{a.e. in } T,$$

$$\dot{x}_0(t) = u(t), \quad x_0(0) = x_1(0) = 0,$$

$$u(t) \in [0, 1] \quad \text{a.e. in } [-q\theta, 1],$$

where $\theta \in (0, 1/q]$ and u is a measurable function mapping $[-q\theta, 1]$ to \mathbf{R} . These problems are expressed in the form we have been considering if we set $\xi = (0, 0)$, $g(x_0, x_1) = x_1$, and

$$f(t, x_0, x_1, u, v, w) = (u, (x_0 - t/2)^2 + h(u, v, w)).$$

In this way, problem $(P_{p,q})$ is posed over $\mathcal{A}(f, \mathcal{U}(p\theta, q\theta))$.

Example 3.3 ($p = 1$). Consider problem $(P_{1,q})$ where, for all $(u, v, w) \in \mathbf{R}^3$,

$$h(u, v, w) = \text{Min} \{ |(u - 1, v - 1, w)|, |(u, v, w - 1)| \}.$$

Let $\mu(t) := \frac{1}{2} \delta_{(1,1,0)} + \frac{1}{2} \delta_{(0,0,1)}$ for all $t \in T$. Clearly, μ is a weakly relaxed control for $(P_{1,q})$ and its corresponding cost is zero. Since the cost cannot be negative, this implies that

$$\text{Min } g(\mathcal{R}(f, \mathcal{M}_w(\theta, q\theta))) = 0.$$

We will show that $\text{Inf } g(\mathcal{R}(f, \mathcal{U}(\theta, q\theta))) > 0$, so that minimizing weakly relaxed controls cannot be approximated by ordinary ones, and so the (weak) extension is not proper. Let (x_0, x_1, u) be any admissible original process for $(P_{1,q})$ and set

$$\hat{u}(t) := (u(t), u(t - \theta), u(t - q\theta)) \quad (t \in T).$$

Suppose we show that almost everywhere in $[0, 1 - (q - 1)\theta]$,

$$\varphi(t) := \sum_{i=0}^{q-1} h(\hat{u}(t + i\theta)) \geq 1.$$

Then we would have

$$q \int_0^1 h(\hat{u}(t)) dt \geq \int_0^{1-(q-1)\theta} \varphi(t) dt \geq 1 - (q - 1)\theta \geq \theta,$$

implying that $x_1(1) \geq \theta/q$ and, hence, $\inf g(\mathcal{R}(f, \mathcal{U}(\theta, q\theta))) \geq \theta/q > 0$. For this purpose, fix $t \in [0, 1 - (q-1)\theta]$ and define, for all $i = -q, 1-q, \dots, q-1$,

$$r_i := u(t + i\theta)$$

and, for all $i = 0, 1, \dots, q-1$, let $\hat{r}_i := h(\hat{u}(t + i\theta))$ and

$$x_i := \begin{cases} 1 & \text{if } \hat{r}_i = |(r_i - 1, r_{i-1} - 1, r_{i-q})|, \\ 0 & \text{if } \hat{r}_i = |(r_i, r_{i-1}, r_{i-q} - 1)|. \end{cases}$$

Suppose first that, for some $i \in \{0, 1, \dots, q-2\}$, $x_i \neq x_{i+1}$. Then

$$\varphi(t) \geq \hat{r}_i + \hat{r}_{i+1} \geq |1 - r_i| + |r_i| \geq 1.$$

So, we may assume that $x_i = x_0$ for all $i \in \{1, 2, \dots, q-1\}$. Then, however,

$$\varphi(t) \geq \hat{r}_0 + \hat{r}_{q-1} \geq |1 - r_{-1}| + |r_{-1}| \geq 1,$$

and the result follows.

Example 3.4 ($p=2$). Consider now problem $(P_{2,q})$ with $h: \mathbf{R}^3 \rightarrow \mathbf{R}$ defined, for all $(u, v, w) \in \mathbf{R}^3$, by

$$h(u, v, w) = \min \{ |(u-1, v, w-1)|, |(u, v-1, w)| \}.$$

A similar argument to the one of the previous example applies if we set $\mu(t) := \frac{1}{2}\delta_{(1,0,1)} + \frac{1}{2}\delta_{(0,1,0)}$ for all $t \in T$. Again, this is a weakly relaxed control for $(P_{2,q})$ and its corresponding cost is zero, so that

$$\min g(\mathcal{R}(f, \mathcal{M}_w(2\theta, q\theta))) = 0.$$

If $(x_0, x_1, u) \in \mathcal{A}(f, \mathcal{U}(2\theta, q\theta))$ and, as before, we set

$$\hat{u}(t) := (u(t), u(t-2\theta), u(t-q\theta)) \quad (t \in T),$$

we prove that $\varphi(t) \geq 1$ almost everywhere in $[0, 1 - (q-1)\theta]$, implying, as above, that

$$\inf g(\mathcal{R}(f, \mathcal{U}(2\theta, q\theta))) \geq \theta/q > 0.$$

Using the same notation for r_i and \hat{r}_i as in Example 3.3, let, for all $i = 0, 1, \dots, q-1$,

$$x_i := \begin{cases} 1 & \text{if } \hat{r}_i = |(r_i - 1, r_{i-2}, r_{i-q} - 1)|, \\ 0 & \text{if } \hat{r}_i = |(r_i, r_{i-2} - 1, r_{i-q})|. \end{cases}$$

Since we are assuming that 2 and q are relatively primes, q is of the form $2r+1$ for some $r \in \mathbf{N}$.

Case 1. $r = 2m-1$ for some $m \in \mathbf{N}$. Assume that $\varphi(t) < 1$. If $x_0 = 1$, then $x_2 = 0$ for, otherwise,

$$\varphi(t) \geq \hat{r}_0 + \hat{r}_2 \geq |1 - r_0| + |r_0| \geq 1.$$

Similarly, we must have

$$x_{4i} = 1 \quad \text{for } i = 0, 1, \dots, m-1,$$

$$x_{4i+2} = 0 \quad \text{for } i = 0, 1, \dots, m-1.$$

Thus, $x_{q-1} = x_{4(m-1)+2} = 0$, and so $x_1 = 1$. Again, this implies that

$$x_{4i+1} = 1 \quad \text{for } i = 0, 1, \dots, m-1,$$

$$x_{4i+3} = 0 \quad \text{for } i = 0, 1, \dots, m-2,$$

and so $x_{q-2} = x_{4(m-1)+1} = 1$. Then, however,

$$\varphi(t) \geq \hat{r}_0 + \hat{r}_{q-2} \geq |1 - r_{-2}| + |r_{-2}| \geq 1.$$

Starting with $x_0 = 0$, a similar argument shows that $\varphi(t) \geq 1$.

Case 2. $r = 2m$ for some $m \in \mathbb{N}$. Assuming again that $\varphi(t) < 1$ we have, if $x_0 = 1$ that

$$x_{4i} = 1 \quad \text{for } i = 0, 1, \dots, m,$$

$$x_{4i+2} = 0 \quad \text{for } i = 0, 1, \dots, m-1,$$

and so $x_{q-1} = x_{4m} = 1$. Thus $x_1 = 0$ and this implies that

$$x_{4i+1} = 0 \quad \text{for } i = 0, 1, \dots, m-1,$$

$$x_{4i+3} = 1 \quad \text{for } i = 0, 1, \dots, m-1.$$

Hence, $x_{q-2} = x_{4(m-1)+3} = 1$, and so

$$\varphi(t) \geq \hat{r}_0 + \hat{r}_{q-2} \geq |1 - r_{-2}| + |r_{-2}| \geq 1.$$

The proof for the case $x_0 = 0$ is similar.

We can proceed as above to show that, for $p = 3$, the function h of Example 3.3 provides a problem for which the infimum cost is strictly greater than the minimum cost of the weakly relaxed version. For $p = 4$ and the function h of Example 3.4, the same conclusion holds.

4. An approximation result. In this section we prove a result that shows some of the difficulties that appear in trying to find a suitable relaxation procedure for a system with noncommensurate delays. Though the results of this section remain valid for systems involving any finite number of constant delays, we analyze, for simplicity, systems with only two delays. The last assertion in the paper does not provide a proper relaxation for systems with noncommensurate delays, but, instead, it says in what sense the closure of the reachable set for a noncommensurate delay problem can be “approximated” in terms of strongly relaxed processes.

Let us denote by \mathcal{K} the class of all nonempty compact subsets of \mathbf{R}^n , and by δ the Hausdorff metric on \mathcal{K} (see [3]). Recall that if for all $x \in \mathbf{R}^n$, $A \in \mathcal{K}$, and $\varepsilon > 0$, we set

$$d(x, A) := \inf \{|x - y| : y \in A\},$$

$$S(A; \varepsilon) := \{x \in \mathbf{R}^n \mid d(x, A) < \varepsilon\},$$

then, for all $A, B \in \mathcal{K}$, the metric δ is given by

$$\delta(A, B) = \inf \{\alpha > 0 \mid A \subset S(B; \alpha) \text{ and } B \subset S(A; \alpha)\}.$$

Let $K(a, b) := \overline{\mathcal{R}(f, \mathcal{U}(a, b))}$ for all $(a, b) \in \mathcal{D}_2 = \{(\theta_1, \theta_2) \in \mathbf{R}^2 \mid 0 < \theta_1 < \theta_2 \leq 1\}$. Our assumptions listed in § 1 imply the existence of an integrable function $\psi : T \rightarrow \mathbf{R}$ such that, for all $(t, r) \in T \times \Omega^3$ and $(x, \mu) \in \mathcal{A}(f, \mathcal{M}(T, \Omega^3))$,

$$|f(t, x(t), r)| \leq \psi(t)$$

(see [2] and [6] for details), and so $K(\mathcal{D}_2) \subset \mathcal{K}$.

If the function $K(\cdot, \cdot)$ mapping \mathcal{D}_2 into the Hausdorff metric space of nonempty compact subsets of \mathbf{R}^n were continuous, we could then, in view of Theorem 3.2, approximate the closure of the reachable set of any noncommensurate delay problem in terms of strongly relaxed trajectories. The continuity of $K(\cdot, \cdot)$ is equivalent to the

statement that, given a point $(a, b) \in \mathcal{D}_2$, a sequence $\{(a_n, b_n)\} \subset \mathcal{D}_2$ converging to (a, b) , and any $\varepsilon > 0$, then, for all n sufficiently large,

$$(4.1) \quad K(a_n, b_n) \subset S(K(a, b); \varepsilon)$$

and

$$(4.2) \quad K(a, b) \subset S(K(a_n, b_n); \varepsilon).$$

The following example shows that (4.1) may not hold.

Example 4.1. Consider the function

$$h(u, v, w) = \text{Min} \{ |(u-1, v-1, w)|, |(u, v, w-1)| \}$$

of Example 3.3 and, for all $n \in \mathbb{N}$, let $u_n \in \mathcal{U}([-1/2, 1], [0, 1])$ alternately take on the values 0 and 1 on successive intervals of length $1/8n$ in $[-1/2, 1]$. Explicitly, for all $n \in \mathbb{N}$ and all $k = -2n, -2n+1, \dots, 4n-1$, let

$$u_n(t) := \begin{cases} 0 & \text{if } t \in [k/4n, (2k+1)/8n), \\ 1 & \text{if } t \in [(2k+1)/8n, (k+1)/4n). \end{cases}$$

Let $a := \frac{1}{4}$, $b := 2a$, and $b_n := b - (1/8n)$ for all $n \in \mathbb{N}$. By construction, we readily verify that, for all $n \in \mathbb{N}$,

$$h(u_n(t), u_n(t-a), u_n(t-b_n)) = 0.$$

On the other hand, applying the inequality

$$q \int_0^1 h(\hat{u}(t)) dt \geq 1 - (q-1)\theta$$

in Example 3.3, we obtain that

$$\int_0^1 h(u(t), u(t-a), u(t-b)) dt \geq \frac{3}{8}$$

for any control $u \in \mathcal{U}([-b, 1], [0, 1])$.

Set $\xi := 0$ and, for all $t \in T$, $x \in \mathbf{R}$ and $(u, v, w) \in \mathbf{R}^3$, let

$$f(t, x, u, v, w) := h(u, v, w).$$

Note that if y_n is the solution of

$$\begin{aligned} \dot{y}(t) &= f(t, y(t), u_n(t), u_n(t-a), u_n(t-b_n)) \quad \text{a.e. in } T, \\ y(0) &= \xi, \end{aligned}$$

then $0 = y_n(1) \in K(a, b_n)$ for all $n \in \mathbb{N}$, but

$$\begin{aligned} d(y_n(1), K(a, b)) &= \text{Inf} \{ |x| : x \in K(a, b) \} \\ &= \text{Inf } \mathcal{R}(f, \mathcal{U}(a, b)) \\ &\geq \frac{3}{8}, \end{aligned}$$

showing that relation (4.1) does not hold.

Though this example may be discouraging, showing that $K(\cdot, \cdot)$ may fail to be continuous, we next prove that this function is lower semicontinuous in the sense that (4.2) does hold.

THEOREM 4.2. The function $\overline{\mathcal{R}(f, \mathcal{U}(\cdot, \cdot))}$, mapping \mathcal{D}_2 into the Hausdorff metric space of nonempty compact subsets of \mathbf{R}^n , is lower semicontinuous.

Proof. Let $(a, b) \in \mathcal{D}_2$ and let $\{(a_n, b_n)\} \subset \mathcal{D}_2$ be any sequence converging to (a, b) . To prove (4.2), let $\varepsilon > 0$ and $p \in K(a, b)$. In view of Lemma 2.1, there exists $u \in \mathcal{U}([-b, 1], \Omega)$ such that $|x(1) - p| < \varepsilon/2$, where x is the solution of

$$\dot{x}(t) = f(t, x(t), \tilde{u}(t)) \quad \text{a.e. in } T, \quad x(0) = \xi,$$

and $\tilde{u}(t) = (u(t), u(t-a), u(t-b))$ for all $t \in T$. Extend the control u to the interval $[-1, 1]$ by assigning some $\omega \in \Omega$ to $u(t)$ for $t \in [-1, -b)$ and, for all $n \in \mathbb{N}$ and $t \in T$, set $\tilde{u}_n(t) = (u(t), u(t-a_n), u(t-b_n))$. Let x_n be the solution of

$$\dot{x}(t) = f(t, x(t), \tilde{u}_n(t)) \quad \text{a.e. in } T, \quad x(0) = \xi.$$

Since $x_n(1) \in K(a_n, b_n)$, the result will follow if we show that $|x_n(1) - x(1)| \rightarrow 0$, $n \rightarrow \infty$ for then, for n sufficiently large, $d(p, K(a_n, b_n)) < \varepsilon$, and (4.2) holds.

To begin with, note that for all $n \in \mathbb{N}$ and $t \in T$,

$$|x_n(t) - x(t)| \leq \int_0^1 z_n(t) dt + \int_0^t \phi(s) |x_n(s) - x(s)| ds,$$

where

$$z_n(t) := |f(t, x(t), \tilde{u}_n(t)) - f(t, x(t), \tilde{u}(t))|,$$

and the integrable function ϕ (see § 1) corresponds to the Lipschitz rank of f with respect to the state. Now, since $u \in \mathcal{U}([-1, 1], \Omega)$, it follows by Lusin's theorem that, for some $A \subset [-1, 1]$ closed, u restricted to A is continuous and $m([-1, 1] - A) < \varepsilon$, where m denotes the Lebesgue measure. Thus, if we set

$$B_1 := \{t \in T \mid t - a \in A\}, \quad B_2 := \{t \in T \mid t - b \in A\},$$

and $B := A \cap B_1 \cap B_2$, the function $\tilde{u}(t)$ is continuous and hence uniformly continuous on B . Since $f(t, x(t), \tilde{u})$ is continuous and hence uniformly continuous on $B \times \Omega^3$, there exists a $\delta \in (0, \varepsilon)$ such that $t, t' \in B$ and $|t - t'| < \delta$ implies that

$$|f(t, x(t), \tilde{u}(t)) - f(t, x(t), \tilde{u}(t'))| < \varepsilon.$$

Note that, since $a_n \rightarrow a$ and $b_n \rightarrow b$, there exists an integer N_0 such that $n > N_0$ implies $|a_n - a| < \delta$ and $|b_n - b| < \delta$. Set

$$B_{1n} := \{t \in T \mid t - a_n \in A\}, \quad B_{2n} := \{t \in T \mid t - b_n \in A\}.$$

If $n > N_0$ and $t \in B \cap B_{1n} \cap B_{2n}$, then $|f(t, x(t), \tilde{u}_n(t)) - f(t, x(t), \tilde{u}(t))| < \varepsilon$.

Now, let $R := |\xi| + \int_0^1 \psi(t) dt$ and

$$M := 2 \cdot \text{Max} \{|f(t, x, r)| : |x| \leq R, r \in \Omega^3 \text{ and } t \in T\}.$$

Then, for $n > N_0$, we have

$$\begin{aligned} \int_0^1 z_n(t) dt &= \int_{T-B \cap B_{1n} \cap B_{2n}} z_n(t) dt + \int_{B \cap B_{1n} \cap B_{2n}} z_n(t) dt \\ &\leq Mm(T - B \cap B_{1n} \cap B_{2n}) + \varepsilon. \end{aligned}$$

Observe that

$$\begin{aligned} (T - B_1) + \{-a\} &= [-a, 1 - a] - A, \\ (T - B_2) + \{-b\} &= [-b, 1 - b] - A, \\ (T - B_{1n}) + \{-a_n\} &= [-a_n, 1 - a_n] - A, \\ (T - B_{2n}) + \{-b_n\} &= [-b_n, 1 - b_n] - A. \end{aligned}$$

Thus,

$$\begin{aligned}
 m(T - B \cap B_{1n} \cap B_{2n}) &= m((T - A) \cup (T - B_1) \cup (T - B_2) \cup (T - B_{1n}) \cup (T - B_{2n})) \\
 &\leq m(T - A) + m(T - B_1) + m(T - B_2) \\
 &\quad + m(T - B_{1n}) + m(T - B_{2n}) \\
 &\leq m(T - A) + m([-a, 1 - a] - A) + m([-b, 1 - b] - A) \\
 &\quad + m([-a_n, 1 - a_n] - A) + m([-b_n, 1 - b_n] - A) \\
 &\leq 5\varepsilon.
 \end{aligned}$$

Therefore, for all $n > N_0$ and $t \in T$,

$$|x_n(t) - x(t)| \leq (1 + 5M)\varepsilon + \int_0^t \phi(s) |x_n(s) - x(s)| ds.$$

Applying Gronwall's inequality, we obtain, for all $n > N_0$ and $t \in T$,

$$|x_n(t) - x(t)| \leq (1 + 5M)\varepsilon \left(1 + c \int_0^t \phi(s) ds \right),$$

where $c = \exp(\int_0^1 \phi(t) dt)$. Thus, $|x_n(1) - x(1)| \rightarrow 0$, $n \rightarrow \infty$ and the result follows.

Combining this result with Theorem 3.2, we obtain the following approximation result for noncommensurate delay problems.

COROLLARY 4.3. *Suppose that we are given $(\theta_1, \theta_2) \in \mathcal{D}_2$. Let $\{(a_n, b_n)\} \subset \mathcal{D}_2$ be any sequence converging to (θ_1, θ_2) such that, for all $n \in \mathbb{N}$, $a_n/b_n = p_n/q_n$, where $1 \leq p_n < q_n$ are relatively primes. For all $n \in \mathbb{N}$, set $\theta_n := a_n/p_n = b_n/q_n$. Then, for any $\varepsilon > 0$ and all n sufficiently large,*

$$\overline{\mathcal{R}(f, \mathcal{U}(\theta_1, \theta_2))} \subset S(\mathcal{R}(\tilde{f}_{p_n, q_n}, \mathcal{M}_{q_n}(\theta_n)); \varepsilon).$$

Acknowledgment. I am grateful to the referee, who made substantial improvements to the exposition of the whole paper and the proof of Theorem 4.2.

REFERENCES

- [1] T. ANDREWS, *An existence theory for optimal control problems with time delays*, Ph. D. thesis, Imperial College, University of London, 1989.
- [2] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.
- [3] K. KURATOWSKI, *Topology*, Vol. 1, Academic Press, New York, 1966.
- [4] J. ROSENBLUETH AND R. B. VINTER, *Relaxation procedures for time delay systems*, J. Math. Anal. Appl., 162 (1991), pp. 542–563.
- [5] J. ROSENBLUETH, *A proper relaxation of optimal control problems*, J. Optim. Theory Appl., to appear.
- [6] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
- [7] ———, *Optimal controls with pseudodelays*, SIAM J. Control Optim., 12 (1974), pp. 286–299.
- [8] ———, *Nonadditively Coupled Delayed Controls*, privately circulated, 1986.

UNMIXED SOLUTIONS OF THE DISCRETE-TIME ALGEBRAIC RICCATI EQUATION*

H. K. WIMMER†

Abstract. The algebraic Riccati equation of the optimal control problem associated with the discrete-time system $x(k+1) = Fx(k) + Gu(k)$ is studied. It is shown that in the case of a controllable system, there exist solutions with prescribed unmixed characteristic polynomial of the corresponding closed-loop matrix. Existence of solutions will also be proved under the weaker condition of modulus-controllability. Maximal solutions are discussed.

Key words. discrete-time algebraic Riccati equation, symplectic pencils, unmixed solutions, modulus-controllability, maximal solution

AMS(MOS) subject classifications. 15A24, 93C55

1. Introduction. A solution X of the continuous-time algebraic Riccati equation (CARE)

$$(1.1) \quad F^*X + XF - XGG^*X - Q = 0$$

is called *unmixed* [11] if the closed-loop matrix $F - GG^*X$ and the matrix $-(F - GG^*X)^*$ have only purely imaginary eigenvalues in common. According to Shayman [11], [12], such solutions share properties of maximal and minimal solutions of the CARE (1.1). In this paper, we are concerned with the algebraic Riccati equation that is the discrete-time counterpart of (1.1). We consider the discrete-time algebraic Riccati equation (DARE)

$$(1.2) \quad \Re(X) = X - F^*XF + F^*XG(I + G^*XG)^{-1}G^*XF - Q = 0,$$

where F , G , Q are complex matrices of sizes $n \times n$, $n \times p$, and $n \times n$, respectively, and Q is positive semidefinite ($Q \geq 0$). Only Hermitian matrices that satisfy (1.2) will be regarded as solutions. We prove an existence and uniqueness result for unmixed solutions of (1.2) and discuss maximal solutions. Our approach is based on the associated symplectic pencil

$$(1.3) \quad M - zL = \begin{pmatrix} F & 0 \\ -Q & I \end{pmatrix} - z \begin{pmatrix} I & \Gamma \\ 0 & F^* \end{pmatrix}, \quad \Gamma = GG^*.$$

Notation. Let $\sigma(F)$ denote the spectrum of F . We write $|\sigma(F)| = 1$, respectively, $|\sigma(F)| \leq 1$ if all eigenvalues of F lie on the unit circle, respectively, in the closed unit disc. A complex number λ is a *characteristic root* of the pencil $M - zL$ if $\det(M - \lambda L) = 0$. Let $g(z) = \prod_{\nu=1}^n (\lambda_\nu - z)$ be a complex polynomial. Put

$$\tilde{g}(z) = \prod_{\nu=1}^n (1 - \bar{\lambda}_\nu z).$$

We call g an *unmixed polynomial* if g and \tilde{g} have only zeros α in common (if any) with $|\alpha| = 1$. In other words, if $g(\lambda) = 0$ and $|\lambda| \neq 1$, $\lambda \neq 0$, then $g(\bar{\lambda}^{-1}) \neq 0$. In particular, g is unmixed if all its roots lie in the closed unit disc. We say that X is an *unmixed*

* Received by the editors August 22, 1990; accepted for publication (in revised form) May 17, 1991. This research was supported by Stimulation Programme of the Commission of the European Communities grant EG SC1/0126-C and by Deutsche Forschungsgemeinschaft grant Kn 164/3-1.

† Mathematisches Institut, Universität Würzburg, D-8700 Würzburg, Germany.

solution of (1.2) if the characteristic polynomial $\det(F_X - zI)$ of its associated closed-loop matrix

$$(1.4) \quad F_X = (I + \Gamma X)^{-1} F = F - G(I + G^* X G)^{-1} G^* X F$$

is unmixed. Given the pencil (1.3), we see later that it is possible to factorize its determinant into

$$(1.5) \quad \det(M - zL) = cg(z)\tilde{g}(z), \quad c \in \mathbb{C},$$

if all unimodular eigenvalues of F are G -controllable. We call (1.5) an *unmixed factorization* if the polynomial g is unmixed.

Let $K = K(F, G) = \text{Im}(G, FG, \dots, F^{n-1}G)$ be the (F, G) -controllable subspace of \mathbb{C}^n . Put $\bar{K} = \mathbb{C}^n / K$. Since the matrix F leaves K invariant, it induces an endomorphism \bar{F} on \bar{K} . Define

$$(1.6) \quad h(z) = \det(zI - \bar{F}).$$

With respect to an appropriate basis of \mathbb{C}^n , the matrices F and G have the form

$$(1.7) \quad F = \begin{pmatrix} F_1 & 0 \\ F_{21} & F_2 \end{pmatrix}, \quad G = \begin{pmatrix} 0 \\ G_2 \end{pmatrix},$$

where the pair (F_2, G_2) is controllable. Then F_1 is a matrix representation of \bar{F} , and we have

$$(1.8) \quad h(z) = \det(zI - F_1).$$

In the case of the CARE (1.1), the pair (F, G) is called sign-controllable (see, e.g., [4]) if the polynomials $h(z)$ and $\bar{h}(-z)$ are coprime or, equivalently, if $\text{rank}(F - \lambda I, G) < n$ implies $\text{rank}(F^* + \bar{\lambda}I, G) = n$. The counterpart to sign-controllability in the case of the DARE (1.2) will be described in the following definition.

DEFINITION 1.1. Let the polynomial h be defined as in (1.6) or (1.8). We call the pair (F, G) *modulus-controllable* if $(h, \tilde{h}) = 1$ or, equivalently, if $|\lambda\mu| = 1$ implies $\text{rank}(F - \lambda I, G) = n$ or $\text{rank}(F - \mu I, G) = n$.

If the pair (F, G) is stabilizable, then we have $|\sigma(F_1)| < 1$; hence (F, G) is modulus-controllable.

The main result of the paper is the following theorem.

THEOREM 1.2. *Let*

$$(1.3)' \quad M - zL = \begin{pmatrix} F - zI & -z\Gamma \\ -Q & I - zF^* \end{pmatrix}, \quad \Gamma = GG^*$$

be the pencil associated to the DARE

$$(1.2)' \quad X - F^* X F + F^* X G (I + G^* X G)^{-1} G^* X F - Q = 0$$

and let

$$(1.4)' \quad F_X = (I + \Gamma X)^{-1} F$$

be the closed-loop matrix corresponding to the solution X . Assume that (F, G) is modulus-controllable and let the polynomial h be defined as in (1.6) or (1.8). Then there exists an unmixed factorization

$$(1.5)' \quad \det(M - zL) = cg(z)\tilde{g}(z)$$

such that

$$(1.9) \quad (h, \tilde{g}) = 1.$$

To each unmixed factorization (1.5)' satisfying (1.9), there exists a unique solution X with $\det(F_X - zI) = g(z)$.

For the CARE (1.1), a result analogous to Theorem 1.2 is available [13]. In the case of the DARE (1.2)', stabilizability of (F, G) is the weakest assumption known to guarantee the existence of a solution [3].

2. Reduction to the controllable case. The following lemma, together with the remarks at the end of this section, should make it clear why the concept of modulus-controllability will play an important role for existence and uniqueness of solutions.

LEMMA 2.1. Assume that F and G are of the form (1.7)

$$F = \begin{pmatrix} F_1 & 0 \\ F_{21} & F_2 \end{pmatrix}, \quad G = \begin{pmatrix} 0 \\ G_2 \end{pmatrix},$$

and put $\Gamma_2 = G_2 G_2^*$ such that $\Gamma = \text{diag}(0, \Gamma_2)$. Let

$$Q = \begin{pmatrix} Q_1 & Q_{12} \\ Q_{12}^* & Q_2 \end{pmatrix}$$

be partitioned accordingly. Define

$$(2.1) \quad M_2 - zL_2 = \begin{pmatrix} F_2 - zI & -z\Gamma_2 \\ -Q_2 & I - zF_2^* \end{pmatrix}$$

and put

$$(2.2) \quad \hat{F}_2 = (I + \Gamma_2 X_2)^{-1} F_2.$$

Then

$$(2.3) \quad \det(M - zL) = \det(F_1 - zI) \det(I - zF_1^*) \det(M_2 - zL_2).$$

A matrix

$$(2.4) \quad X = \begin{pmatrix} X_1 & X_{12} \\ X_{12}^* & X_2 \end{pmatrix}$$

is a solution of (1.2) if and only if it consists of blocks that satisfy the following set of equations:

$$(2.5a) \quad \Re_2(X_2) = X_2 - F_2^* X_2 F_2 + F_2^* X_2 G_2 (I + G_2^* X_2 G_2)^{-1} G_2^* X_2 F_2 - Q_2 = 0,$$

$$(2.5b) \quad X_{12} - F_1^* X_{12} \hat{F}_2 = B,$$

$$(2.5c) \quad X_1 - F_1^* X_1 F_1 = C,$$

where

$$(2.6) \quad B = Q_{12} + F_{21}^* X_2 \hat{F}_2$$

and

$$(2.7) \quad C = F_{21}^* X_{12}^* F_1 + (F_1^* X_{12} + F_{21}^* X_2) [-\Gamma_2 X_{12}^* (I + \Gamma_2 X_2)^{-1} F_1 + \hat{F}_2] + Q_1.$$

For the closed-loop matrix associated to (2.4), we have

$$(2.8) \quad \det(zI - F_X) = \det(zI - F_1) \det(zI - \hat{F}_2).$$

Proof. The factorization (2.3) is obvious. Using the matrix identity $I - G(I + G^* X G)^{-1} G^* X = (I + \Gamma X)^{-1}$ we can write (1.2) as

$$(2.9) \quad \Re(X) = X - F^* X (I + \Gamma X)^{-1} F - Q = X - F^* X F_X - Q = 0.$$

With the matrices

$$F_X = (I + \Gamma X)^{-1} F = \begin{pmatrix} F_1 & 0 \\ -\Gamma_2 X_{12} (I + \Gamma_2 X_2)^{-1} F_1 + \hat{F}_2 & \hat{F}_2 \end{pmatrix}$$

and

$$F^* X = \begin{pmatrix} F_1^* X_1 + F_{21}^* X_{12} & F_1^* X_{12} + F_{21}^* X_2 \\ F_2^* X_{12} & F_2^* X_2 \end{pmatrix}$$

at hand, it is not difficult to verify that (2.9) is equivalent to (2.5) \square

Suppose that a basis of \mathbb{C}^n is chosen such that F and G are as in (1.7) and the pair

$$(2.10) \quad (F_2, G_2) \text{ is controllable.}$$

Starting from (2.5a), i.e., from the Riccati equation $\mathfrak{R}_2(X_2) = 0$ that fulfils hypothesis (2.10), a solution X of (1.2) can be obtained in two steps by solving the linear matrix equations (2.5b) and (2.5c). It will be seen that unique solvability of (2.5b) is equivalent to $(\tilde{h}, g) = 1$. Given the blocks X_2 and X_{12} , there exists a unique solution X_1 of (2.5c) if $1 \notin \sigma(F_1^*)\sigma(F_1)$, which is equivalent to the condition $(h, \tilde{h}) = 1$ of modulus-controllability of (F, G) .

3. Basic facts of the DARE and the associated symplectic pencil. The pencil $M - zL$ given by (1.3) plays a crucial role in the study of the DARE. Most of the statements of the following lemma are well known. We refer to [9], [3], [8], and [15].

LEMMA 3.1. (i) *Let R and S be nonsingular complex $2n \times 2n$ matrices such that*

$$(3.1) \quad (M - zL)R = S \begin{pmatrix} \Lambda - zI & -zD \\ 0 & I - z\Lambda^* \end{pmatrix}, \quad D = D^*.$$

Let

$$(3.2) \quad R = \begin{pmatrix} R_1 & \cdot \\ R_{21} & \cdot \end{pmatrix}$$

be partitioned into $n \times n$ blocks. If R_1 is nonsingular and $X = R_{21}R_1^{-1}$ is Hermitian, then $I + \Gamma X$ (and hence $I + G^* X G$) is nonsingular [15] and X is a solution of (1.2). Suppose that $\det(\Lambda - zI) = g(z)$ holds; then (1.5) holds, and $F_X = R_1 \Lambda R_1^{-1}$ implies $\det(F_X - zI) = g(z)$.

(ii) *Let X be a solution of (1.2). Then*

$$(3.3) \quad (M - zL) \begin{pmatrix} I & 0 \\ X & I \end{pmatrix} = \begin{pmatrix} I + \Gamma X & 0 \\ F^* X & I \end{pmatrix} \begin{pmatrix} F_X - zI & -zD \\ 0 & I - zF_X^* \end{pmatrix},$$

where $D = D^* = (I + \Gamma X)^{-1} \Gamma = G(I + G^* X G)^{-1} G^*$. Furthermore,

$$(3.4) \quad \det(M - zL) = c \det(F_X - zI) \det(I - zF_X^*).$$

The main feature of $M - zL$ is the relation

$$(3.5) \quad MJM^* = LJL^*,$$

where J is given by $J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}$. Property (3.5) characterizes symplectic pencils. Elementary divisors of $M - zL$ corresponding to a characteristic root λ appear in pairs $(\lambda - z)^\nu$ and $(1 - z\bar{\lambda})^\nu$ if $\lambda \neq 0$ and $|\lambda| \neq 1$. A pairing also exists between elementary divisors of the form z^ν and infinite elementary divisors [9]. Hence an unmixed factorization of

$\det (M - zL)$ exists if and only if all unimodular characteristic roots have even algebraic multiplicity. If $\det (M - zL) \neq 0$ whenever $|\alpha| = 1$, then (1.5) is an unmixed factorization if and only if $(g, \tilde{g}) = 1$. In that case, there exist matrices R and S such that (3.1) holds with $D = 0$ and $g(z)$ is the characteristic polynomial of Λ .

4. Characteristic roots on the unit circle. In this section, we see that the proof of Theorem 1.2 can be reduced to the case of a pencil $M - zL$ without unimodular characteristic roots.

Notation. The generalized eigenspace corresponding to an eigenvalue λ of F will be denoted by $E_\lambda(F)$, i.e., $E_\lambda(F) = \text{Ker} (F - \lambda I)^n$. Let

$$V = V(F, Q) = \text{Ker} \begin{pmatrix} Q \\ QF \\ \vdots \\ QF^{n-1} \end{pmatrix}$$

be the weakly unobservable subspace of \mathbb{C}^n . For $\lambda \in \sigma(F)$, put

$$V_\lambda = V_\lambda(F, Q) = E_\lambda(F) \cap V.$$

Since V is invariant under F , we have

$$V = \oplus [E_\lambda(F) \cap V], \quad \lambda \in \sigma(F),$$

and we could define V_λ as the maximal F -invariant subspace of \mathbb{C}^n contained in $E_\lambda(F) \cap \text{Ker } Q$.

In [3] the existence of unimodular characteristic roots of $M - zL$ is related to the rank conditions (4.2) and (4.3) below.

LEMMA 4.1. Assume that $|\alpha| = 1$. Then we have

$$(4.1) \quad \det (M - \alpha L) = 0$$

if and only if

$$(4.2) \quad \text{rank} (F - \alpha I, \Gamma) < n$$

or

$$(4.3) \quad \text{rank} \begin{pmatrix} F - \alpha I \\ Q \end{pmatrix} < n.$$

Proof. Let $w \in \mathbb{C}^{2n}$, $w \neq 0$, be such that

$$(4.4) \quad (M - \alpha L)w = 0.$$

Put $w^T = (w_1^T, w_2^T)$. It is easy to see that (4.4) and $|\alpha| = 1$ imply $-w_1^*(I - \alpha F^*) - w_2^* \Gamma = 0$, and $-Qw_1 + (I - \alpha F^*)w_2 = 0$, which yields

$$-w_1^*(I - \alpha F^*)w_2 - w_2^* \Gamma w_2 - w_1^* Qw_1 + w_1^*(I - \alpha F^*)w_2 = 0.$$

From $\Gamma \geq 0$, $Q \geq 0$, and $w_2^* \Gamma w_2 + w_1^* Qw_1 = 0$, we obtain $\Gamma w_2 = Qw_1 = 0$. Hence

$$(4.5) \quad \begin{pmatrix} F^* - \bar{\alpha} I \\ \Gamma \end{pmatrix} w_2 = 0$$

and

$$(4.6) \quad \begin{pmatrix} F - \alpha I \\ Q \end{pmatrix} w_1 = 0.$$

From $w \neq 0$ follows (4.2) or (4.3). That (4.2) or (4.3) implies (4.1) is obvious, since (4.5) and (4.6) yield $(0, w_2^*)(M - zL) = 0$ and

$$(M - zL) \begin{pmatrix} w_1 \\ 0 \end{pmatrix} = 0,$$

respectively. \square

LEMMA 4.2. Let $U = (U_1, U_2)$ be a nonsingular $n \times n$ matrix with $U_i \in \mathbb{C}^{n \times n_i}$, $i = 1, 2$. Then the columns of U_1 form a basis of V_α if and only if

$$(4.7) \quad U^{-1}FU = \begin{pmatrix} A_1 & A_{12} \\ 0 & A_2 \end{pmatrix}, \quad U^*QU = \text{diag}(0, Q_2),$$

$$(4.8) \quad \sigma(A_1) = \{\alpha\},$$

and A_2, Q_2 are of size $n_2 \times n_2$ and

$$(4.9) \quad \text{rank} \begin{pmatrix} A_2 - \alpha I \\ Q_2 \end{pmatrix} = n_2.$$

Proof. Let \mathfrak{U} denote the set of F -invariant subspaces contained in $E_\alpha(F) \cap \text{Ker } Q$. Then (4.7) and (4.8) are equivalent to $\text{span } U_1 \in \mathfrak{U}$. Suppose now that

$$(4.10) \quad \text{rank} \begin{pmatrix} A_2 - \alpha I \\ Q_2 \end{pmatrix} < n_2.$$

Then we have $(A_2 - \alpha I)w = 0$, $Q_2w = 0$, for some $w \neq 0$. Hence $\text{span}(U_1, U_2w) \in \mathfrak{U}$ and $\text{span } U_1$ is not maximal in \mathfrak{U} . To prove the converse, suppose that $\text{span}(U_1, y) \in \mathfrak{U}$ for some $y \notin U$. We can assume that $y \in \text{span } U_2$ such that $y = U_2w$, $w \neq 0$. Then $Qy = 0$ and, therefore, $Q_2w = 0$. From $Fy = FU \begin{pmatrix} 0 \\ w \end{pmatrix} = U_2A_2w$ follows $Fy \in \text{span } U_2 \cap \text{span}(U_1, y) \cap E_\alpha(F)$. Hence we have $Fy = \alpha y$, which implies $A_2w = \alpha w$. We have found that

$$\begin{pmatrix} A_2 - \alpha I \\ Q_2 \end{pmatrix} w = 0, \quad w \neq 0,$$

which yields (4.10). \square

The notation U^{-*} below is used for the matrix $(U^*)^{-1}$.

LEMMA 4.3. Consider an eigenvalue α of F such that $|\alpha| = 1$, $E_\alpha(F) \cap \text{Ker } Q \neq 0$, and

$$(4.11) \quad \text{rank}(F - \alpha I, \Gamma) = n.$$

Let $U = (U_1, U_2)$ be a nonsingular matrix that transforms F and Q as in (4.7) and (4.8). Let

$$(4.12) \quad U^{-1}\Gamma U^{-*} = \begin{pmatrix} \cdot & \cdot \\ \cdot & \Gamma_2 \end{pmatrix}$$

be partitioned, conforming to (4.7), and put

$$(4.13) \quad M_2 - zL_2 = \begin{pmatrix} A_2 & 0 \\ -Q_2 & I \end{pmatrix} - z \begin{pmatrix} I & \Gamma_2 \\ 0 & A_2^* \end{pmatrix}.$$

Then

$$(4.14) \quad \text{span } U_1 = V_\alpha$$

if and only if

$$(4.15) \quad \det(M_2 - \alpha L_2) \neq 0.$$

Proof. We know from Lemma 4.1 that $\det(M_2 - \alpha L_2) = 0$ holds if and only if the matrices $(A_2 - \alpha I, \Gamma_2)$ and

$$\begin{pmatrix} A_2 - \alpha I \\ Q_2 \end{pmatrix}$$

do not both have maximal rank. Now (4.11) and $\Gamma \geq 0$ imply that $\text{rank}(A_2 - \alpha I, \Gamma_2) = n_2$. Hence (4.15) is equivalent to (4.9), which, according to the previous lemma, is equivalent to (4.14). \square

The following result shows that a unimodular characteristic root α yields a subspace V_α , which lies in the kernel of each solution X .

LEMMA 4.4. *Let α be a characteristic root of $M - zL$ with $|\alpha| = 1$, which satisfies (4.11).*

(i) *For each solution X of the DARE (1.2), we have $V_\alpha \subseteq \text{Ker } X$. Furthermore, $E_\alpha(F_X) = V_\alpha$ and $F_X = F$ on V_α .*

(ii) *Let $U = (U_1, U_2)$ be nonsingular such that $\text{span } U_1 = V_\alpha$, let the pencil $M_2 - zL_2$ be given as in (4.13), and assume that α is not a characteristic root of $M_2 - zL_2$. Put*

$$(4.16) \quad U^{-1}G = \begin{pmatrix} G_1 \\ G_2 \end{pmatrix}.$$

A matrix X is a solution of (1.2) if and only if

$$(4.17) \quad X = U^{-*} \text{diag}(0, X_2) U^{-1},$$

and X_2 is a solution of the DARE

$$(4.18) \quad X_2 - A_2^* X_2 A_2 - A_2^* X_2 G_2 (I + G_2^* X_2 G_2)^{-1} G_2^* X_2 A_2 - Q_2 = 0.$$

Proof. (i) According to [14], condition (4.11) yields

$$E_\alpha(F_X) \subseteq E_\alpha(F) \cap \text{Ker } Q \cap \text{Ker } X,$$

and $F_X = F$ on $E_\alpha(F_X)$. Hence $E_\alpha(F_X) \in \mathcal{U}$. Put $k = \dim E_\alpha(F_X)$. Then (3.4) implies that

$$\det(M - zL) = (z - \alpha)^{2k} b(z), \quad b(\alpha) \neq 0.$$

Let $U = (U_1, U_2)$ be a matrix as in Lemma 4.2 such that $\text{span } U_1 = V_\alpha$. Recall $n_1 = \dim V_\alpha$. From

$$(4.19) \quad \det(M - zL) = \det(F_1 - zI) \det(I - zF_1^*) \det(M_2 - zL_2)$$

and Lemma 4.3, we obtain $\det(M - zL) = (z - \alpha)^{2n_1} f(z)$, $f(\alpha) \neq 0$. Hence $k = n_1$ and $E_\alpha(F_X) = V_\alpha$.

(ii) It is easy to verify that each X_2 coming from (4.18) yields a solution X given by (4.17). It is not obvious, however, that under hypotheses (4.11) all solutions of (1.2) should be of the form (4.17). We know from part (i) that $V_\alpha = \text{span } U_1 \subseteq \text{Ker } X$. Hence $U^* X U = \text{diag}(0, X_2)$, and X_2 is a solution of (4.18). \square

5. Auxiliary results, proof of Theorem 1.2. A matrix R in (3.1) and (3.2) yields a solution of (1.2) only if R_1^{-1} exists. To prove nonsingularity of R_1 , we use a result on the discrete-time Lyapunov matrix equation.

LEMMA 5.1. *Let Λ and P be complex $n \times n$ matrices such that*

$$(5.1) \quad 1 \notin \sigma(\Lambda^*) \sigma(\Lambda)$$

and $P \geq 0$. If Y is a solution of

$$(5.2) \quad Y - \Lambda^* Y \Lambda = P,$$

then $Y = Y^*$ and

$$\text{Ker } Y = V(\Lambda, P) = \text{Ker} \begin{pmatrix} P \\ P\Lambda \\ \vdots \\ P\Lambda^{n-1} \end{pmatrix}.$$

Proof. Choose a basis of \mathbb{C}^n such that

$$V(\Lambda, P) = \left\{ \begin{pmatrix} x_1 \\ 0 \end{pmatrix}, x_1 \in \mathbb{C}^{n_1} \right\}.$$

Then

$$\Lambda = \begin{pmatrix} \Lambda_1 & \Lambda_{12} \\ 0 & \Lambda_2 \end{pmatrix}, \quad P = \text{diag}(0, P_2),$$

and the pair (Λ_2^*, P_2) is controllable. Because of $1 \notin \sigma(\Lambda_2^*)\sigma(\Lambda_2)$, the equation $Y_2 - \Lambda_2^* Y_2 \Lambda_2 = P_2$, $P_2 \geq 0$ has a unique Hermitian solution Y_2 that is nonsingular [16]. Then $Y = \text{diag}(0, Y_2)$ is a solution of (5.2) with $\text{Ker } Y = V(\Lambda, P)$ and, because of (5.1), the solution is unique. \square

The uniqueness statement of Theorem 1.2 will follow from a result of Willems (see [10, p. 197]).

LEMMA 5.2. *Let X and W be two solutions of (1.2); then*

$$(5.3) \quad X - W = F_X^*(X - W)F_W.$$

Proof. Since [10] seems to be the only reference for (5.3), we include a proof. Recall (2.9) and note that $F = (I + \Gamma W)F_W$ and $F^* = F_X^*(I + X\Gamma)$. Then

$$\begin{aligned} \Re(X) - \Re(W) &= X - W - (F_W^*XF - F^*WF_W) \\ &= X - W - [F_X^*X(I + \Gamma W)F_W - F_X^*(I + X\Gamma)WF_W] \\ &= X - W - F_X^*(X - W)F_W, \end{aligned}$$

which yields (5.3). \square

After a reduction to the controllable case that was carried out in § 2, we are able to discard unimodular characteristic roots of $M - zL$. The following lemma will justify such a simplification. Since previous results are extended from V_α to $\oplus\{V_\alpha, |\alpha| = 1\}$, we refer to matrices and equations of the preceding section, making the provision that (4.8) is to be replaced by $|\sigma(A_1)| = 1$.

LEMMA 5.3. *Assume that*

$$(5.4) \quad |\alpha| = 1 \text{ implies } \text{rank}(F - \alpha I, \Gamma) = n.$$

(i) *There exists a nonsingular matrix U such that (4.7) holds with $|\sigma(A_1)| = 1$ and such that the pencil $M_2 - zL_2$ given by (4.13) and (4.12) has no unimodular characteristic roots.*

(ii) *Put $f(z) = \det(A_1 - zI)$. Then $\det(M - zL) = cg(z)\tilde{g}(z)$ is an unmixed factorization if and only if*

$$(5.5) \quad g(z) = f(z)b(z)$$

and

$$(5.6) \quad \det(M_2 - zL_2) = cb(z)\tilde{b}(z), \quad (b, \tilde{b}) = 1.$$

(iii) A matrix X is a solution of (1.2) if and only if it is of the form (4.17), where X_2 is a solution of the Riccati equation given by (4.18) and (4.16).

(iv) A solution X is unmixed with

$$(5.7) \quad \det(F_X - zI) = g(z) = f(z)b(z)$$

if and only if the matrix X_2 of (4.17) is an unmixed solution of (4.18) such that the closed loop matrix

$$(5.8) \quad \Phi_2 = (I + \Gamma_2 X_2)^{-1} A_2$$

satisfies

$$(5.9) \quad \det(\Phi_2 - zI) = b(z).$$

Proof. Parts (i) and (iii) are immediate consequences of Lemmas 4.2 and 4.4. Note that $|\sigma(A_1)| = 1$ is equivalent to $f = \gamma \tilde{f}$, $\gamma \in \mathbb{C}$. Hence (5.5) and (5.6) follow from (4.19). For a solution X of the form (4.17), we have

$$F_X = U \begin{pmatrix} A_1 & \Phi_{12} \\ 0 & \Phi_2 \end{pmatrix} U^{-1},$$

with Φ_2 as in (5.8). Hence $\det(F_X - zI) = f(z) \det(\Phi_2 - zI)$, and (5.7) is equivalent to (5.9). \square

THEOREM 5.4. Assume that all unimodular eigenvalues α of F are G -controllable, i.e., that condition (5.4) holds. If $\det(M - zL) = cg(z)\tilde{g}(z)$ is an unmixed factorization, then there is at most one solution X of (1.2) such that $\det(F_X - zI) = g(z)$.

Proof. From the preceding lemma we know that the proof involves only a pencil $M_2 - zL_2$ without unimodular characteristic roots. Hence it suffices to prove uniqueness under the assumption $(g, \tilde{g}) = 1$. Suppose that X and W are two solutions such that

$$\det(F_X - zI) = \det(F_W - zI) = g(z).$$

Then $1 \notin \sigma(F_X^*)\sigma(F_W)$, and $\Delta = 0$ is the only solution of $\Delta - F_X^* \Delta F_W = 0$. Thus, according to Lemma 5.2, we have $X - W = 0$. \square

THEOREM 5.5. Suppose that (F, G) is controllable. Then there exists an unmixed factorization of $\det(M - zL)$. To each unmixed factorization $\det(M - zL) = cg(z)\tilde{g}(z)$, there exists a unique solution X such that

$$(5.10) \quad \det(F_X - zI) = g(z).$$

Proof. The fact that the controllability hypothesis (5.4) of Lemma 5.3 holds allows us to work with a pencil $M - zL$ that has no characteristic roots of modulus 1 and, accordingly, to proceed under the assumption $(g, \tilde{g}) = 1$. In that case, there exist nonsingular matrices R and S such that

$$(5.11) \quad (M - zL)R = S \begin{pmatrix} \Lambda - zI & 0 \\ 0 & I - z\Lambda^* \end{pmatrix}$$

and $\det(\Lambda - zI) = g(z)$. Let R be partitioned as in (3.2). Then

$$(5.12) \quad \text{rank} \begin{pmatrix} R_1 \\ R_{21} \end{pmatrix} = n.$$

To obtain a solution X in the form $X = R_{21}R_1^{-1}$, we must make sure that R_1 is nonsingular. Put $Y = R_{21}^*R_1$. We want to first show that

$$(5.13) \quad \text{Ker } Y \subseteq \text{Ker } R_{21}$$

holds. Since $R_1 x = 0$ implies $Yx = 0$ and (5.13) yields $R_{21}x = 0$, we would obtain $x = 0$ from (5.12). Hence as soon as we have established (5.13), we know that R_1 is nonsingular.

The subsequent argument that yields the discrete-time Lyapunov equation (5.16) can be found in [3]. From (5.11) follows

$$\begin{pmatrix} F & 0 \\ -Q & I \end{pmatrix} \begin{pmatrix} R_1 \\ R_{21} \end{pmatrix} = \begin{pmatrix} I & \Gamma \\ 0 & F^* \end{pmatrix} \begin{pmatrix} R_1 \\ R_{21} \end{pmatrix} \Lambda,$$

which is equivalent to the pair of equations

$$(5.14) \quad FR_1 = R_1 \Lambda + \Gamma R_{21} \Lambda$$

and

$$(5.15) \quad -R_1^* Q + R_{21}^* = \Lambda^* R_{21}^* F.$$

Multiplying (5.14) from the left by $\Lambda^* R_{21}^*$ and (5.15) from the right by R_1 are steps that lead to

$$(5.16) \quad Y - \Lambda^* Y \Lambda = \Lambda^* R_{21}^* \Gamma R_{21} \Lambda + R_1^* Q R_1 = P.$$

Since $(g, \tilde{g}) = 1$ is equivalent to (5.1), it follows from Lemma 5.1 that $\text{Ker } Y = V(\Lambda, P)$. Hence $\text{Ker } Y$ is a Λ -invariant subspace spanned by chains of eigenvectors and generalized eigenvectors of Λ , like x_1, \dots, x_k , which satisfy $\Lambda x_i = \lambda x_i + x_{i-1}$, $i = 1, \dots, k$, $x_0 = 0$, $x_1 \neq 0$, and $Px_i = 0$. Induction will show that for such a chain, we have

$$(5.17) \quad x_j \in \text{Ker } R_{21}$$

for $j = 0, 1, \dots, k$. Assume that (5.17) holds for $j = i - 1$. Then $Px_i = 0$, and $\Gamma \geq 0$, $Q \geq 0$ imply that

$$(5.18) \quad \Gamma R_{21} \Lambda x_i = 0$$

and

$$(5.19) \quad QR_1 x_i = 0.$$

From (5.14) and (5.15) we obtain

$$(5.20) \quad R_{21} x_i = F^* R_{21} \Lambda x_i.$$

In the case where $\lambda = 0$, we find that $R_{21} x_i = F^* R_{21} x_{i-1}$ and the induction hypotheses yield $R_{21} x_i = 0$. In the case where $\lambda \neq 0$, we conclude from (5.18) and (5.15) that $\Gamma R_{21} \Lambda x_i = 0$ and $R_{21} x_i = F^* R_{21} \lambda x_i$. Hence

$$(R_{21} x_i)^* (\bar{\lambda}^{-1} I - F, \Gamma) = 0.$$

In this case, controllability of (F, Γ) implies $R_{21} x_i = 0$.

To show that $X = R_{21} R_1^{-1}$ is Hermitian, note that because of (5.1) the matrix $Y = R_{21}^* R_1$ is a unique, and hence Hermitian, solution of (5.2). Therefore $Y = R_1^* R_{21}$ and $X = R_1^{-1} Y R_1^{-1}$. Hence, X is also Hermitian.

From (5.14) we obtain $F = (I + \Gamma X) R_1 \Lambda R_1^{-1}$ and

$$(5.21) \quad F_X = R_1 \Lambda R_1^{-1}.$$

Lemma 3.1(i) tells us that we have found a solution X of (1.2) with the desired property (5.10). By Theorem 5.4 such a solution is unique, which completes the proof. \square

From (5.16) and (5.21) follows the equation

$$(5.22) \quad X - F_X^* X F_X = F_X^* X \Gamma X F_X + Q,$$

which leads to inertia results for Riccati equations (see [1]).

Proof of Theorem 1.2. We now perform the construction described at the end of § 2. We assume that F and G are given as in (1.7)

$$F = \begin{pmatrix} F_1 & 0 \\ F_{21} & F_2 \end{pmatrix}, \quad G = \begin{pmatrix} 0 \\ G_2 \end{pmatrix},$$

such that (F_2, G_2) is controllable, and $h(z) = \det(F_1 - zI)$. Modulus-controllability of (F, G) is equivalent to $(h, \tilde{h}) = 1$. Let $\det(M - zL) = cg(z)\tilde{g}(z)$ be an unmixed factorization that satisfies $(\tilde{h}, g) = 1$. Then (2.3) implies $h|g$. Put $f = g/h$ and let $M_2 - zL_2$ be the pencil (2.1). Then $\det(M_2 - zL_2) = cf(z)\tilde{f}(z)$ is an unmixed factorization. Since (2.5a) is a Riccati equation where the pair (F_2, G_2) is controllable, we know from Theorem 5.5 that $\Re_2(X_2) = 0$ has a unique solution X_2 such that $\det(\hat{F}_2 - zI) = f(z)$, where \hat{F}_2 is given by (2.2). The solution X_2 enters into the definition of B in (2.6). From $(\tilde{h}, g) = 1$ follows $(\tilde{h}, f) = 1$ or, equivalently, $1 \notin \sigma(F_1^*)\sigma(\hat{F}_2)$. Hence (2.5b) has a unique solution X_{12} . Given X_1 and X_{12} , the matrix C in (2.7) is well defined. Now consider (2.5c). Modulus-controllability amounts to $1 \notin \sigma(F_1^*)\sigma(F_1)$. Hence (2.5c) determines X_1 uniquely. The block matrix (2.4) is a solution of (1.2). From (2.8) we obtain $\det(F_X - zI) = h(z)f(z) = g(z)$, and X is the only solution with that property. \square

6. Maximal solutions. It is known that (1.2) has a solution if the pair (F, G) is stabilizable [3]. In that case [7], there exists a maximal solution X with the properties $X \geq 0$ and

$$(6.1) \quad |\sigma(F_X)| \leq 1.$$

In this section we focus on property (6.1) and its relation with maximality. As a stabilizable pair, (F, G) is necessarily modulus-controllable; the following result is a special case of Theorem 5.5. The existence statement in the subsequent theorem can be found in [3].

THEOREM 6.1. *If (F, Γ) is stabilizable, then there exists a unique solution X of (1.2) such that $|\sigma(F_X)| \leq 1$.*

A solution X is called *maximal* if $X - W \geq 0$ holds for all solutions W of (1.2). We see that (6.1) is equivalent to maximality of X , provided that the standing assumption (5.4) holds. Two auxiliary results will be needed.

LEMMA 6.2 (see [2]). *Let X and W be two solutions of (1.2). Then $\Delta = X - W$ satisfies the equation*

$$(6.2) \quad \Delta - F_X^* \Delta F_X = F_X^* \Delta G (I + G^* W G)^{-1} G^* \Delta F_X.$$

LEMMA 6.3. *If X and W are two solutions of (1.2), then*

$$(6.3) \quad \ln(I + G^* X G) = \ln(I + G^* W G).$$

Proof. Relation (6.3) appears in [6] where (1.2) is approached by factorization results of matrices of rational functions under the hypotheses that (F, G) is controllable and $|\sigma(F)| < 1$. Here we use the pencil $M - zL$. It is easy to verify that (3.3) implies that

$$(6.4) \quad \begin{pmatrix} I & 0 \end{pmatrix} (M - zL)^{-1} \begin{pmatrix} 0 \\ I \end{pmatrix} = z(F_X - zI)^{-1} (I + \Gamma X)^{-1} \Gamma (I - zF_X^*)^{-1}.$$

Note that $(I + \Gamma X)^{-1}\Gamma = G(I + G^*XG)^{-1}G^*$. Consider (6.4) and the corresponding expression for the solution W and take $z = \alpha$, where $|\alpha| = 1$ and $\det(M - zL) \neq 0$. If the symbol \sim denotes congruence, then

$$(6.5) \quad G(I + G^*XG)^{-1}G \sim G(I + G^*WG)^{-1}G.$$

It is not difficult to show that (6.5) implies (6.3). \square

THEOREM 6.4. *Assume that $\text{rank}(F - \alpha I, \Gamma) = n$ for all α with $|\alpha| = 1$. If X is a solution of (1.2) that satisfies $|\sigma(F_X)| \leq 1$, then X is a maximal solution.*

Proof. According to Lemma 5.3, each solution X of (1.2) is of the form $X = U^{-*} \text{diag}(0, X_2)U^{-1}$, where X_2 is a solution of a Riccati equation whose associated pencil $M_2 - zL_2$ has no unimodular characteristic roots. Hence we can assume for the proof that X is a solution with the property

$$(6.6) \quad |\sigma(F_X)| < 1.$$

It is a known application of (5.22) that (6.6) implies $X \geq 0$. Therefore $I + G^*XG > 0$, and by the preceding lemma we have $I + G^*WG > 0$ for all solutions W . Put $\Delta = X - W$ and define $S = F_X^* \Delta G(I + G^*WG)^{-1}G^* \Delta F_X$. Then $\Delta - F_X^* \Delta F_X = S$ is (6.2). From (6.6) and $S \geq 0$ follows $\Delta \geq 0$; hence X is a maximal solution. \square

Acknowledgment. The author thanks Dr. C. Scherer for valuable comments.

REFERENCES

- [1] S. BITTANTI, P. BOLZERN, AND P. COLANERI, *Inertia theorems for Lyapunov and Riccati equations—an updated view*, in Proc. SIAM Conference on Linear Algebra in Signals, Systems, and Control, Boston, 1986, pp. 11–35, B. N. Datta et al., eds. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1988.
- [2] S. W. CHAN, G. C. GOODWIN, AND K. S. SIN, *Convergence properties of the Riccati difference equation in optimal filtering of nonstabilizable systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 110–118.
- [3] C. E. DE SOUZA, M. R. GEVERS, AND G. C. GOODWIN, *Riccati equations in optimal filtering of nonstabilizable systems having singular state transition matrices*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 831–838.
- [4] L. E. FAIBUSOVICH, *Algebraic Riccati equation and symplectic algebra*, Internat. J. Control, 43 (1986), pp. 781–792.
- [5] V. KUČERA, *The discrete Riccati equation of optimal control*, Kybernetika, 8 (1972), pp. 430–447.
- [6] P. LANCASTER, A. C. M. RAN, AND L. RODMAN, *Hermitian solutions of the discrete algebraic Riccati equation*, Internat. J. Control, 44 (1986), pp. 777–802.
- [7] ———, *An existence and monotonicity theorem for the discrete algebraic matrix Riccati equation*, Linear and Multilinear Algebra, 20 (1987), pp. 353–361.
- [8] V. MEHRMANN, *The Linear-Quadratic Control Problem: Theory and Numerical Algorithms*, Habilitationsschrift, Universität Bielefeld, Bielefeld, Germany, 1987.
- [9] T. PAPPAS, A. J. LAUB, AND N. R. SANDELL, *On the numerical solution of the discrete-time algebraic Riccati equation*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 631–641.
- [10] G. PICCI, *Elementi di Elaborazione Statistica del Segnale*, CLEUP Editore, Padova, Italy, 1986.
- [11] M. A. SHAYMAN, *Geometry of the algebraic Riccati equation, Part I*, SIAM J. Control Optim., 21 (1983), pp. 375–394.
- [12] ———, *Geometry of the algebraic Riccati equation, Part II*, SIAM J. Control Optim., 21 (1983), pp. 395–409.
- [13] H. K. WIMMER, *The algebraic Riccati equation: conditions for the existence and uniqueness of solutions*, Linear Algebra Appl., 58 (1984), pp. 441–452.
- [14] ———, *Strong solutions of the discrete-time algebraic Riccati equation*, Systems Control Lett., 13 (1989), pp. 455–457.
- [15] ———, *Normal forms of symplectic pencils and the discrete-time algebraic Riccati equation*, Linear Algebra Appl., 147 (1991), pp. 411–440.
- [16] H. K. WIMMER AND A. D. ZIEBUR, *Remarks on inertia theorems for matrices*, Czechoslovak. Math. J., 25 (1975), pp. 556–561.

ON THE EXISTENCE OF CONTROL LYAPUNOV FUNCTIONS: GENERALIZATIONS OF VIDYASAGAR'S THEOREM ON NONLINEAR STABILIZATION*

JOHN TSINIAS†

Abstract. In this paper the well-known Vidyasagar's theorem concerning the feedback stabilizability problem for interconnected control systems is generalized. In particular, sufficient conditions are provided for the existence of control Lyapunov functions that, according to the results of Artstein, Sontag, and Tsinias, guarantees asymptotic stabilization by means of a feedback law that is smooth, except possibly at the equilibrium at which it is wished to stabilize the system.

Key words. control Lyapunov functions, state feedback stabilizability

AMS(MOS) subject classification. 93D15

1. Introduction. The paper deals with the state feedback stabilization problem of nonlinear systems at a specified equilibrium. Sufficient conditions for local and global stabilization are presented for a wide class of systems that are affine in the control. The results of the paper generalize the well-known theorems of Vidyasagar on asymptotic stabilization [27] and considerably improve those developed in our recent papers, [25] and [26].

Our purpose is to provide sufficient conditions for the existence of suitable control Lyapunov functions that according to [3], [17], and [23]–[26] guarantee stabilization by means of a feedback law that is smooth (C^∞) except possibly at the equilibrium.

We consider systems of the form

$$(1.1) \quad \dot{x} = F(x) + G(x)u,$$

where R^n and R^l are the state and the input space, respectively, and zero $0 \in R^n$ is an equilibrium for the uncontrolled term F ; i.e., $F(0) = 0$. We assume that the mappings F and G have the form

$$(1.2) \quad F = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \quad \text{and} \quad G_i = \begin{pmatrix} 0 \\ g_i \end{pmatrix}, \quad i = 1, \dots, l,$$

where $f_1: R^n \rightarrow R^{n_1}$, $f_2: R^n \rightarrow R^{n_2}$, and $g_i: R^n \rightarrow R^{n_2}$, $n_1 + n_2 = n$ are Lipschitz continuous. According to decomposition (1.2), system (1.1) is written as

$$(1.3) \quad \begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} f_1(x) \\ f_2(x) \end{pmatrix} + \begin{pmatrix} 0 \\ g(x)u \end{pmatrix}, \quad x = (x'_1, x'_2)' \in R^{n_1} \times R^{n_2},$$

where $'$ stands for transpose.

As pointed out in [18] and [25], the regularity assumptions for the stabilizing feedback near the equilibrium play an important role in the theory that has been developed. Linear, smooth, almost smooth, and piecewise analytic feedback controllers have been used in [1]–[9], [11]–[13], and [15]–[27], and the various types of regularity requirements lead to many different notions of stabilization.

* Received by the editors September 10, 1990; accepted for publication (in revised form) April 15, 1991.

† National Technical University, Department of Mathematics, Zografou Campus, 157 73 Athens, Greece.

We say that (1.1) is (*globally*) *asymptotically stabilizable* if there exists a feedback law $u = k(x)$ that is smooth for $x \neq 0$ and such that zero $0 \in R^n$ is (*globally*) asymptotically stable for the resulting closed-loop system

$$(1.4) \quad \dot{x} = (F + Gk)(x).$$

System (1.1) is *globally exponentially stabilizable* if, furthermore, there exist positive constants $\alpha > 0$ and $\beta > 0$ such that

$$\|x(t, x_0)\| \leq \alpha \exp(-\beta t) \|x_0\|, \quad \forall t \geq 0, \quad x_0 \in R^n,$$

where $x(t, x_0)$ denotes the trajectory of (1.4) of time t starting at x_0 , and $\|\cdot\|$ is the usual Euclidean norm.

We say that system (1.1) satisfies the *Lyapunov condition* (lc), if there exist a neighborhood N of $0 \in R^n$ and a real function $\Phi: N \rightarrow R$, which is at least continuously differentiable on N , is positive definite, i.e., $\Phi(0) = 0$ and $\Phi(x) > 0$ for $x \in N \setminus \{0\}$, and such that for any $x \in N \setminus \{0\}$ it follows that

$$(1.5) \quad G(\Phi)(x) \stackrel{\text{def}}{=} (D\Phi G_1, \dots, D\Phi G_l)(x) = 0 \Rightarrow F(\Phi)(x) < 0,$$

where $D\Phi$ denotes the derivative of Φ . A continuously differentiable real function Φ is called a *control Lyapunov function* (clf) if it is positive definite and satisfies condition (1.5).

We say that the above clf Φ satisfies the *small control property* if, furthermore, there exists a nonnegative real function $m: N \rightarrow R^+$ such that $m(x) \rightarrow 0$ as $x \rightarrow 0$, and for every $x \in N \setminus \{0\}$ there exists a vector $u \in R^l$ satisfying the following inequalities: $\|u\| < m(x)$ and $F(\Phi)(x) + G(\Phi)(x)u < 0$.

A control Lyapunov function $\Phi: R^n \rightarrow R^+$ is called a *global* clf if it satisfies (1.5) for every nonzero $x \in R^n$, and, furthermore, it is *uniformly unbounded on R^n* , namely, $\Phi(x) \rightarrow +\infty$ as $\|x\| \rightarrow +\infty$.

In [3] it is shown that the lc is a necessary and sufficient condition for asymptotic stabilization. Furthermore, there exists a global clf if and only if the system is globally asymptotically stabilizable. The proposed stabilizing feedback law will be continuous at the origin if and only if the corresponding clf satisfies the small control property. Versions and generalizations of this theorem are also provided in [17] and [22]–[26]. In particular, Sontag [17] provides an explicit formula for the stabilizing feedback.

Let us now focus our attention on the nonlinear case described by (1.3). It will be useful here to recall the precise statement of Vidyasagar's theorem on stabilization. Suppose that zero $0 \in R^{n_1}$ is asymptotically stable with respect to $\dot{x}_1 = f_1(x_1, 0)$, f_2 and g are independent of x_1 , and the system $\dot{x}_2 = f_2(x_2) + g(x_2)u$ is stabilizable at $x_2 = 0$ by means of a Lipschitz continuous map $k: R^{n_2} \rightarrow R^l$. Then Theorem 3.1 in [27] asserts that the same feedback $u = k(x_2)$ also asymptotically stabilizes (locally) the overall system (1.3) at zero $0 \in R^n$. Its proof follows by an interesting Lyapunov-like approach based on the comparison principle. Some additional assumptions (see Theorems 3.2 and 3.3 in [27]) guarantee global and exponential stabilizability. We note that the same approach leads to further generalizations. In particular, suppose that there exists a *continuously differentiable* map $\phi: R^{n_1} \rightarrow R^{n_2}$, $\phi(0) = 0$ such that $0 \in R^{n_1}$ is asymptotically stable with respect to

$$(1.6) \quad \dot{x}_1 = f_1(x_1, \phi(x_1)),$$

and there exists a Lipschitz continuous feedback law $u = k(x)$ such that the set $M_\phi \stackrel{\text{def}}{=} \{x \in R^n : x_2 = \phi(x_1)\}$ is asymptotically stable with respect to (1.4). In particular,

assume that $x_2(t, x_0) - \phi(x_1(t, x_0)) \rightarrow 0$, as $t \rightarrow t + \infty$ for any x_0 near zero. It follows that M_ϕ must be an invariant manifold with respect to (1.4), or, equivalently,

$$(1.7) \quad D\phi f_1(x_1, \phi(x_1)) = (f_2 + gk)(x_1, \phi(x_1)), \quad x_1 \in R^{n_1}.$$

Moreover, the same feedback $u = k(x)$ asymptotically stabilizes (1.3) at the origin. The proof of this statement follows if we use the nonlinear change of coordinates $y_1 = x_1, y_2 = x_2 - \phi(x_1)$. Then system (1.3) becomes

$$(1.8a) \quad \dot{y}_1 = f_1(y_1, y_2 + \phi(y_1)),$$

$$(1.8b) \quad \dot{y}_2 = (f_2 + gk)(y_1, y_2 + \phi(y_1)) - D\phi f_1((y_1, \phi(y_1))).$$

Note that $0 \in R^{n_1}$ is asymptotically stable with respect to (1.8a) with $y_2 = 0$, whereas (1.7) guarantees that $0 \in R^{n_2}$ is an equilibrium for (1.8b). In addition, $0 \in R^{n_2}$ is asymptotically stable with respect to (1.8b) uniformly on y_1 . Then we can apply for the transformed system (1.8) exactly the same arguments as those in the proof of Theorem 3.1 in [27] to establish that, indeed, $0 \in R^{n_1} \times R^{n_2}$ is asymptotically stable with respect to (1.8) and therefore to the original (1.3).

The above approach is based on the differentiability of ϕ and on the existence of a mapping $k: R^n \rightarrow R^l$ satisfying the positive invariance property (1.7). Obviously, if one of these assumptions is dropped, the previous approach does not work. However, further interesting generalizations are feasible.

Our main purpose is to show that if ϕ is continuous and if certain Lyapunov-like assumptions are imposed for (1.3) that are weaker from those previously discussed, then (1.3) is (locally) asymptotically stabilizable.

In particular, assume that if ϕ is continuous, there exists a nonnegative continuously differentiable map $W: R^n \rightarrow R$ such that $W(x) = 0$ if and only if $x \in M_\phi$ and condition (1.5) holds for every $x \notin M_\phi$ near zero, and with W instead of Φ . Then, in Theorem 2.2, using a quite different approach from that developed in [27], we prove that the previous assumptions guarantee the existence of a clf, and so system (1.3) is (locally) asymptotically stabilizable. In particular, we show that there exists a Lyapunov function $V: R^{n_1} \rightarrow R^+$ of zero with respect to (1.6) such that the function

$$(1.9) \quad \Phi(x) = V(x_1) + W(x)$$

is a clf for (1.3).

A sufficient condition that depends directly on the stability behavior of system (1.6) and the overall system (1.3) near $0 \in R^{n_1}$ and M_ϕ , respectively, is provided in Theorem 2.4.

A second aim of the paper is to derive sufficient conditions for the existence of a global clf guaranteeing *global* asymptotic stabilization for (1.3) (Theorems 3.2 and 3.5). These conditions are different from those proposed in [27]. Among other things, we prove that if certain additional assumptions are imposed for the function W , then there is a uniformly unbounded Lyapunov function $V: R^{n_1} \rightarrow R^+$ of zero with respect to (1.6) such that the map (1.9) is a global clf for (1.3). If the assumptions of Theorems 3.2 and 3.5 are strengthened further, then in Theorem 4.2 we prove that there exists a Lyapunov function $V: R^{n_1} \rightarrow R^+$ of zero with respect to (1.6) such that for any sufficiently small constant $q > 0$ the map $\Phi(x) = qV(x_1) + W(x)$ is a global clf, and, furthermore, (1.3) is globally *exponentially* stabilizable.

Finally, to illustrate the theory we develop and, in addition, to show how our method is applicable to a number of cases not covered by the earlier work of Vidyasagar and others, we study three numerical examples.

2. Local stabilization. We now give the main sufficient conditions for the existence of control Lyapunov functions for system (1.3).

Suppose that there exist neighborhoods $N_1 \subset R^{n_1}$ and $N_2 \subset R^{n_2}$ of zero, a continuous map $\phi: N_1 \rightarrow R^{n_2}$, $\phi(0) = 0$, and a continuously differentiable function $W: N \stackrel{\text{def}}{=} N_1 \times N_2 \rightarrow R$ satisfying the following conditions:

(A1) The origin $0 \in R^n$ is an asymptotically stable equilibrium with respect to (1.6).

(A2) The function W satisfies the following properties:

$$(2.1a) \quad W(x) = 0, \text{ if and only } x \in M_\phi \stackrel{\text{def}}{=} \{x \in N: x_2 = \phi(x_1), x_1 \in N_1\},$$

$$(2.1b) \quad W(x) > 0, \text{ otherwise,}$$

and, furthermore, for each $x \in N \setminus M_\phi$, the following holds:

$$(2.2) \quad \left(\frac{\partial W}{\partial x_2} g \right)(x) \stackrel{\text{def}}{=} \left(\frac{\partial W}{\partial x_2} g_1, \dots, \frac{\partial W}{\partial x_2} g_l \right)(x) = 0 \Rightarrow F(W)(x) < 0.$$

We also need the following additional condition, which is a special case of (A2):

(A2)' There is a real function $W: N \rightarrow R$ as in (A2), which, in addition, satisfies the following property. There exists a nonnegative real function $e: R^n \rightarrow R^+$ such that $e(x) \rightarrow 0$ as

$$d(x, M_\phi) \stackrel{\text{def}}{=} \inf \{ \|x - y\|, y \in M_\phi \} \rightarrow 0$$

and, for any $x \in N \setminus M_\phi$, a vector $r \in R^l$ can be found satisfying the following inequalities:

$$(2.3) \quad \|r\| < e(x),$$

$$(2.4) \quad F(W)(x) + \left(\frac{\partial W}{\partial x_2} g \right)(x) r < 0.$$

To state and prove our main theorem on local stabilizability, we need the following result.

LEMMA 2.1. (i) Suppose that system (1.3) satisfies condition (A2). Then there exists a map $r: N \rightarrow R^l$, which is smooth on $N \setminus M_\phi$ and such that

$$(2.5) \quad \begin{aligned} r(x) &= 0, \quad \forall x \in M_\phi, \\ \left(F(W) + \frac{\partial W}{\partial x_2} g r \right)(x) &< 0, \quad \forall x \in N \setminus M_\phi. \end{aligned}$$

(ii) Furthermore, if system (1.3) satisfies (A2)', then there exists a map $r: N \rightarrow R^l$ as above, which, in addition, is continuous on N . In particular, $r(x) \rightarrow 0$ as $d(x, M_\phi) \rightarrow 0$.

The proof of the previous lemma follows by partition of unity arguments similar to those given in [3] or [24].

The following theorem generalizes Vidyasagar's theorem on local stabilization. It also considerably improves the results of our recent papers [25] and [26].

THEOREM 2.2. Suppose that system (1.3) satisfies conditions (A1) and (A2). Then there exists a Lyapunov function $\tilde{V}: R^{n_1} \rightarrow R^+$ of zero $0 \in R^{n_1}$ with respect to (1.6) such that the function

$$(2.6) \quad \Phi(x) = \tilde{V}(x_1) + W(x),$$

where W is defined in (A2), is a clf, and so system (1.3) is asymptotically stabilizable. Moreover, if (A1) and (A2)' are fulfilled, then the corresponding clf satisfies the small control property, and, therefore, system (1.3) is asymptotically stabilizable by means of a feedback law that is continuous at the origin (and smooth for $x \neq 0$ near zero).

Proof. Since, by our assumption (A1), zero $0 \in R^n$ is asymptotically stable with respect to (1.6) and the map $x_1 \rightarrow f(x_1, \phi(x_1))$ is continuous, the converse stability theorem of Kurzweil [14] asserts that there exists a smooth Lyapunov function $x_1 \rightarrow V(x_1)$ of zero with respect to (1.6); namely, V is positive definite and satisfies the inequality

$$(2.7) \quad DV(x_1)f_1(x_1, \phi(x_1)) < 0, \quad \forall x_1 \neq 0 \text{ near zero.}$$

Without loss of generality, we assume that N and N_1 are compact and that (2.7) is satisfied on the region $N_1 \setminus \{0\}$. We define

$$M = \{x \in N: DV(x_1)f_1(x_1, x_2) < 0\}.$$

Note that because of (2.7) the set

$$M_\phi^0 = \{x \in N \setminus \{0\}: x_2 = \phi(x_1)\}$$

is contained to M . According to assumption (A2) and Lemma 2.1, there exists a map $r: R^n \rightarrow R^l$, such that $r(x) = 0$ for $x \in M_\phi$, r is smooth on $N \setminus M_\phi$ and satisfies

$$(2.8) \quad E(x) \stackrel{\text{def}}{=} (-F(W) - G(W)r)(x) > 0$$

for every $x \in N \setminus M_\phi$. For each $x_1 \in N_1$ we define

$$a_1(x_1) = \max \{|f_1(V)(x_1, x_2)|, (x'_1, x'_2)' \in N\}.$$

Obviously, a_1 is continuous and nonnegative definite on N_1 . For any $x_1 \neq 0$, consider a closed sphere $B_{\rho_{x_1}}$ of radius ρ_{x_1} centered at $(x_1, \phi(x_1))$, which is contained to M . We define $M' = \bigcup_{x_1 \in N_1 \setminus \{0\}} B_{(1/2)\rho_{x_1}}$ and

$$a_2(x_1) = \inf \{E(x_1, x_2), (x'_1, x'_2)' \in N \setminus M'\}, \quad x_1 \neq 0.$$

Since E is continuous on $N \setminus M'$ and $M_\phi^0 \subset M'$, it follows that a_2 is continuous on $N_1 \setminus \{0\}$ and, furthermore, because of (2.8), a_2 is strictly positive in this region. Therefore there exists a (piecewise linear) continuous real function $b: R^+ \rightarrow R^+$, which is strictly increasing and satisfies

$$(2.9) \quad b(\|x_1\|)a_1(x_1) < a_2(x_1), \quad \forall x_1 \in N_1 \setminus \{0\}.$$

Also, let $a: R^+ \rightarrow R^+$ be a nonnegative strictly increasing continuous function such that

$$(2.10) \quad a(\|x_1\|) \geq V(x_1), \quad \forall x_1 \in N_1.$$

Finally, we define

$$(2.11) \quad \tilde{V}(x_1) = \int_0^{V(x_1)} b(a^{-1}(r)) dr.$$

Obviously, \tilde{V} is positive definite and continuously differentiable on N_1 . Moreover, since a^{-1} is strictly increasing, $b(a^{-1}(V(x_1))) > 0$ for any $x_1 \neq 0$ and so, by (2.7),

$$D\tilde{V}(x_1)f_1(x_1, \phi(x_1)) = DV(x_1)f_1(x_1, \phi(x_1))b(a^{-1}(V(x_1))) < 0;$$

hence \tilde{V} is a Lyapunov function with respect to (1.6). Next, we show that the Lie derivative $(F + Gr)(\Phi)(x)$ is strictly negative for $x \neq 0$, where Φ , r , and \tilde{V} are defined in (2.6), (2.8), and (2.11), respectively. Indeed, we evaluate

$$(2.12) \quad \begin{aligned} (F + Gr)(\Phi)(x) &= D\tilde{V}(x_1)f_1(x) + DW(x)(F + Gr)(x) \\ &= DV(x_1)f_1(x)b(a^{-1}(V(x_1))) - E(x). \end{aligned}$$

For each $x \in M$, we have

$$x_1 \neq 0, \quad b(a^{-1}(V(x_1))) > 0, \quad DV(x_1)f_1(x) < 0, \quad E(x) \geq 0,$$

and, therefore, by (2.12)

$$(2.13) \quad (F + Gr)(\Phi)|_{x \in M} < 0.$$

For $x \in N \setminus M'$ such that $x_1 \neq 0$, it follows by (2.9) and (2.10) that

$$\begin{aligned} DV(x_1)f_1(x)b(a^{-1}(V(x_1))) &< |DV(x_1)f_1(x)|b(\|x_1\|) \\ &\leq a_1(x_1)b(\|x_1\|) < a_2(x_1) \leq E(x), \end{aligned}$$

and so by (2.12) we get

$$(2.14) \quad (F + Gr)(\Phi)|_{x \in N \setminus M', x_1 \neq 0} < 0.$$

Finally, for $x \neq 0$ with $x_1 = 0$, it follows that $x \notin M_\phi$. Therefore $DV(0)f_1(x) = 0$, whereas, by (2.8), $E(x) > 0$. Consequently, by (2.12), we get

$$(2.15) \quad (F + Gr)(\Phi)|_{x_1=0, x \neq 0} < 0.$$

From (2.1), (2.7), and (2.13)–(2.15), it follows that $F(\Phi)(x) < 0$ for every nonzero $x \in N$ such that $G(\Phi)(x) = ((\partial W / \partial x_2)g)(x) = 0$, and so (1.5) is fulfilled. Moreover, since Φ is continuously differentiable and positive definite, we conclude that Φ is a clf, and so (1.3) is asymptotically stabilizable. Suppose now that (A2)' is satisfied. Then, by Lemma 2.1, there is a continuous function $r: N \rightarrow R^l$ such that $r(x) \rightarrow 0$ as $d(x, M_\phi) \rightarrow 0$. The same procedure as above and the fact that $m(x) \stackrel{\text{def}}{=} \|r(x)\| \rightarrow 0$ as $x \rightarrow 0$, guarantees that Φ is a clf that satisfies the small control property. Therefore, in that case, the system is asymptotically stabilizable by means of a feedback law that is continuous at the origin (and smooth for $x \neq 0$ near zero). \square

Remark 2.3. Suppose that (A1) is fulfilled and let us, in addition, assume that (2.4) holds, with r being a Lipschitz continuous map. Then, similar to the proof of Theorem 2.2., we can show that the Lie derivative $(F + Gr)(\Phi)(x)$, where Φ is defined in (2.6), is strictly negative for $x \neq 0$ near zero. The latter, in conjunction with the fact that r is Lipschitz continuous, implies that $0 \in R^n$ is (locally) asymptotically stable with respect to the closed-loop system $\dot{x} = (F + Gr)(x)$.

According to Theorem 2.2, to find a clf it suffices to determine an appropriate nonnegative function W satisfying condition (A2) and then to check the stability behavior of the subsystem $\dot{x}_1 = f_1(x_1, \phi(x_1))$, $x_1 \in R^{n_1}$ at zero. Of course, the previous methodology is useful in several cases, since it considerably simplifies the stability analysis. On the other hand, it presents a theoretical disadvantage because of the presence of the function W , which, in general, cannot be easily determined. The following theorem generalizes Theorem 2.2 in [25]. It provides a sufficient condition for the existence of the function W , which depends directly on the stability behavior of the overall system (1.3) near the set M_ϕ .

THEOREM 2.4. Assume that system (1.3) satisfies condition (A1) and that there exist Lipschitz continuous mappings $p_0: R^n \rightarrow R$ and $p: R^n \rightarrow R^l$ such that p_0 is nonnegative definite and the set $M_\phi = \{x \in R^n: x_2 = \phi(x_1)\}$ is asymptotically stable with respect to

$$(2.16) \quad \dot{x} = (p_0F + Gp)(x).$$

Then system (1.3) satisfies condition (A2), and so it is asymptotically stabilizable.

Proof. Since the set M_ϕ is asymptotically stable with respect to (2.16), it follows by the converse stability theorem of Wilson [28], [29] that there is a smooth real function $W: R^n \rightarrow R$ such that $W(x) = 0$ for $x \in M$, $W(x) > 0$ and $(p_0F + Gp)(W)(x) < 0$

for $x \notin M_\phi$. The latter inequality implies that for any $x \notin M_\phi$ with $G(W)(x) = ((\partial W / \partial x_2)g)(x) = 0$, it holds that $(p_0 F(W))(x) < 0$, and since p_0 is nonnegative definite we get $F(W)(x) < 0$. Therefore (A2) is fulfilled, and so, by Theorem 2.2, system (1.3) is asymptotically stabilizable.

Remark 2.5. It must be pointed out that conditions (A1) and (A2) are also necessary for asymptotic stabilization, provided that M_ϕ is *positively invariant* with respect to the closed-loop system (1.4). Indeed, if zero $0 \in R^n$ is asymptotically stable with respect to (1.4), then $0 \in R^{n_1}$ is also asymptotically stable with respect to (1.6), which is the restriction of (1.4) to M_ϕ , and so (A1) is fulfilled. Finally, assume that k is Lipschitz continuous. Then since $0 \in R^n$ is asymptotically stable with respect to (1.4), M_ϕ has the same property, and so we can apply Theorem 2.4 with $p_0 = 1$ and $p = k$ to establish that (A2) is satisfied, with W being any smooth Lyapunov function of M_ϕ with respect to (2.16).

3. Global stabilization. Next, we give the main sufficient conditions for the existence of a global clf for case (1.3).

Suppose that there exist a continuous map $\phi: R^{n_1} \rightarrow R^{n_2}$, $\phi(0) = 0$ such that $0 \in R^{n_1}$ is globally asymptotically stable with respect to (1.6), and a continuously differentiable function $W: R^n \rightarrow R$ satisfying the following condition:

(B) The function W satisfies (2.1) and (2.2) (with $N = R^n$) and, furthermore,

$$(3.1) \quad W(x) \rightarrow +\infty \quad \text{as} \quad d(x, M_\phi) \rightarrow +\infty.$$

Moreover, there exists a continuously differentiable Lyapunov function V of zero $0 \in R^{n_1}$ with respect to (1.6) and a constant $q_0 > 0$ such that V is uniformly unbounded on R^{n_1} , and for any $x \notin M_\phi$ with $((\partial W / \partial x_2)g)(x) = 0$ a positive constant $q > 0$ can be found satisfying the following inequality:

$$(3.2) \quad F(W)(x) < \min \{0, -qf_1(V)(x)\}.$$

In particular, for sufficiently large x the previous inequality holds for some $q > q_0$.

The following additional condition is a special case of (B):

(B)' There exist real functions W, V , a constant q_0 as in (B), and a function e as in (A2)' such that for any $x \neq 0$ there is a vector $r \in R^l$ with $\|r\| < e(x)$ and a constant $q > 0$ ($q > q_0$ for sufficiently large x) satisfying the following inequality:

$$F(W)(x) + \left(\frac{\partial W}{\partial x_2} g \right)(x) r < \min \{0, -qf_1(V)(x)\}.$$

To prove the main theorem on the global stabilizability, we need to establish the following lemma, which consists of a slight generalization of Lemma 2.1.

LEMMA 3.1. Suppose that there exist continuously differentiable functions $V: R^{n_1} \rightarrow R^+$ and $W: R^n \rightarrow R^+$ satisfying condition (B). Then there exist a strictly increasing continuous function $b: R^+ \rightarrow R^+$ and a map $r: R^n \rightarrow R^l$ with $r(x) = 0$ for $x \in M_\phi$ such that r is smooth on $R^n \setminus M_\phi$ and satisfies the following inequalities:

$$(3.3a) \quad \left(F(W) + \frac{\partial W}{\partial x_2} g \right)(x) < 0, \quad \forall x \in R^n \setminus M_\phi,$$

$$(3.3b) \quad b(\|x_1\|)f_1(V)(x) \leftarrow \left(F(W) + \frac{\partial W}{\partial x_2} g \right)(x),$$

whenever $x \in R^n \setminus M_\phi$ such that $f_1(V)(x) \geq 0$.

The map r will be continuous on R^n if, in addition, we assume that (B) is fulfilled.

Proof. According to condition (B), for every $x \in R^n \setminus M_\phi$, there is a strictly positive constant q ($q > q_0$ for sufficiently large x) and a vector $r \in R^l$ such that

$$(3.4) \quad F(W)(x) + \left(\frac{\partial W}{\partial x_2} g \right)(x) r < \min \{0, -q f_1(V)(x)\}.$$

Indeed, for $x \notin M_\phi$ such that $((\partial W / \partial x_2)g)(x) \neq 0$ and, for any constant $q > q_0$, a vector $r \in R^l$ can be determined such that (3.4) is satisfied. For $x \notin M_\phi$ such that $((\partial W / \partial x_2)g)(x) = 0$, condition (3.4) is an immediate consequence of (3.2). Since $F(W)$, $G(W)$, and $f_1(V)$ are continuous, it follows by (3.4) that for each $z \in R^n \setminus M_\phi$ there exists a closed ball B_{ρ_z} of radius $\rho_z < 1$ centered at z such that (3.4) holds for any $x \in B_{\rho_z}$ and for suitable constants $\bar{q} = q(z) > 0$ ($q(z) > q_0$ for sufficiently large x) and $r = r(z) \in R^l$. Then there is a partition $\{B_i, \psi_i\}$ where $B_i = B_{\rho_{z_i}}$ is locally finite, the union of the interiors of B_i covers $R^n \setminus M_\phi$, $\psi_i = \psi_i(x)$ is a smooth real function supported on B_i with $\psi_i \geq 0$, and $\sum \psi_i = 1$. We define $r_s(x) = \sum \psi_i(x) r(z_i)$, $q_s(x) = \sum \psi_i(x) q(z_i)$ for $x \in R^n \setminus M_\phi$, and $r_s(x) = 0$ for $x \in M_\phi$. Then, since $\sum \psi_i = 1$ and $\{B_i\}$ is locally finite, the mappings r_s and q_s are smooth on $R^n \setminus M_\phi$, and q_s is strictly positive in this region. We can also easily establish that the following holds:

$$(3.5) \quad \left(F(W) + \frac{\partial W}{\partial x_2} g r_s \right)(x) < \min \{0, -q_s(x) f_1(V)(x)\}, \quad \forall x \notin M_\phi,$$

whereas $q_s(x) > q_0$ for sufficiently large x . Moreover, the sets $\{x \in R^n : f_1(V)(x) \geq 0\}$ and $M_\phi \setminus \{0\}$ are disjoint. Therefore there exists a piecewise linear continuous function $b: R^+ \rightarrow R^+$ that is strictly increasing and satisfies $b(\|x_1\|) \leq b(\|x\|) < q_s(x)$ for any $x \in R^n \setminus M_\phi$ with $f_1(V)(x) \geq 0$. The latter inequality, in conjunction with (3.5), implies (3.3). The rest of the proof follows by using similar arguments as before and is left to the reader. \square

The following theorem generalizes Theorem 2.2, of the present paper and asserts that if (B) is satisfied then there exists a global clf.

THEOREM 3.2. *Suppose that system (1.3) satisfies condition (B). Then there exists a uniformly unbounded Lyapunov function $\tilde{V}: R^n \rightarrow R^+$ of zero $0 \in R^n$ with respect to (1.6) such that the function*

$$(3.6) \quad \Phi(x) = \tilde{V}(x_1) + W(x),$$

where W is defined in (B), is a global clf, and so system (1.3) is globally asymptotically stabilizable. The corresponding feedback can be constructed to also be continuous at the origin if we further assume that (B)' is fulfilled.

Proof. According to Lemma 3.1, there exist a map $r: R^n \rightarrow R^l$ with $r(x) = 0$ for $x \in M_\phi$, which is continuous on $R^n \setminus M_\phi$, and a nondecreasing continuous function $b: R^+ \rightarrow R^+$ such that (3.3) is satisfied. Consider next the Lyapunov function V defined in assumption (B) and let $a: R^+ \rightarrow R^+$, $a(0) = 0$ be any strictly increasing continuous function that satisfies

$$(3.7) \quad a(\|x_1\|) \geq V(x_1), \quad \forall x_1 \in R^{n_1}.$$

Finally, consider the function \tilde{V} as defined in (2.11) of Theorem 2.2, where a and b are the real mappings defined as before. Next, we show that the positive definite function Φ , which is defined by (3.6), satisfies condition (1.5). Indeed, for any $x \in M = \{x \in R^n : f_1(V)(x) < 0\}$ it follows by (3.3a) that

$$\left(F(W) + \frac{\partial W}{\partial x_2} g r \right)(x) \leq 0$$

and, similar to the proof of Theorem 2.2, we get $(F + Gr)(\Phi)|_{x \in M} < 0$. For each nonzero $x \notin M$, we have $f_1(V)(x) \geq 0$ and so by (3.3a) and (3.3b) of Lemma 3.1 and (3.7), it follows that

$$\begin{aligned}(F + Gr)(\Phi)(x) &= b(a^{-1}(V(x_1)))f_1(V)(x) + \left(F(W) + \frac{\partial W}{\partial x_2} gr\right)(x) \\ &\leq b(\|x_1\|)f_1(V)(x) + \left(F(W) + \frac{\partial W}{\partial x_2} gr\right)(x) < 0.\end{aligned}$$

We conclude that for each $x \in R^n \setminus \{0\}$ with $G(\Phi)(x) = 0$ it follows that $F(\Phi)(x) < 0$, and so (1.5) is fulfilled for every nonzero $x \in R^n$. Finally, we show that Φ is uniformly unbounded on R^n . Indeed, consider a sequence $\{x_n = (x'_{1n}, x'_{2n})' \in R^n\}$ with $\|x_n\| \rightarrow +\infty$. Suppose first that $\|x_{1n}\| \rightarrow +\infty$. Then since V is uniformly unbounded on R^n and the map $r \rightarrow b(a^{-1}(r))$, $r \in R^+$ is strictly increasing, it follows that for every sufficiently large index k there is an integer $n_0 > k$ such that

$$\begin{aligned}\tilde{V}(x_{1n}) &= \tilde{V}(x_{1k}) + \int_{V(x_{1k})}^{V(x_{1n})} b(a^{-1}(r)) dr \\ &\geq \tilde{V}(x_{1k}) + (V(x_{1n}) - V(x_{1k}))b(a^{-1}(V(x_{1k}))), \quad \forall n > n_0.\end{aligned}$$

Therefore $\tilde{V}(x_{1n}) \rightarrow +\infty$ as $n \rightarrow +\infty$ and so

$$(3.8) \quad \Phi(x_n) = \tilde{V}(x_{1n}) + W(x_n) \rightarrow +\infty.$$

Assume now that $\{x_{1n}\}$ is bounded. Then $\|x_{2n}\| \rightarrow +\infty$ and, consequently,

$$(3.9) \quad d(x_n, M_\phi) \rightarrow +\infty.$$

Condition (3.9) in conjunction with (3.1) implies (3.8), and so Φ is uniformly unbounded on R^n . The result is that Φ is a global clf. The same procedure as above shows that, in addition, Φ satisfies the small control property, provided that (B)' is fulfilled. \square

If the smoothness requirements are relaxed, we can provide an explicit formula for the stabilizing feedback laws for the case of Theorems 2.2 and 3.2. Indeed, suppose that condition (B) is satisfied. Then, according to Sontag's theorem in [17] and our previous Theorem 3.2, the following feedback law

$$k(x) = \begin{cases} 0, & \text{for } x = 0, \\ -\sigma(a(x), b(x))b(x), & \end{cases}$$

where

$$\sigma(a, b) = \begin{cases} 0, & \text{for } b = 0 \text{ and } a < 0, \\ \frac{a + \sqrt{a^2 + \|b\|^2}}{\|b\|^2}, & \text{otherwise} \end{cases}$$

and

$$a \stackrel{\text{def}}{=} F(\Phi) = f_1(\tilde{V}) + F(W),$$

$$b \stackrel{\text{def}}{=} G(\Phi) = \frac{\partial W}{\partial x_2} g$$

is continuous on $R^n \setminus \{0\}$ and globally asymptotically stabilizes (1.3) at the origin. Furthermore, the map $w \rightarrow k(\cdot)w$ minimizes the quadratic cost

$$I = \int_0^{+\infty} (w^2(t) + \|u(t)\|^2) dt$$

subject to the following one-dimensional system parameterized by $x \in R^n$:

$$\dot{w} = \{F(\Phi)(x)\}w + \{G(\Phi)(x)\}u, \quad w \in R.$$

An immediate consequence of Theorems 2.4 and 3.2 is the following proposition.

PROPOSITION 3.3 (see [25]). *Suppose that $0 \in R^{n_1}$ is globally asymptotically stable with respect to (1.6) and that one of the following assumptions holds:*

- (i) *There exists a positive definite real function $W: R^n \rightarrow R$ that satisfies (2.1), (2.2), (3.1), and, furthermore,*

$$(3.10) \quad \left(\frac{\partial W}{\partial x_2} g \right)(x) \neq 0, \quad \forall x \notin M_\phi;$$

- (ii) *There exists a Lipschitz continuous map $p: R^n \rightarrow R$ such that M_ϕ is globally asymptotically stable with respect to $\dot{x} = (Gp)(x)$.*

Then system (1.3) is globally asymptotically stabilizable.

Remark 3.4. Suppose that $n_2 = l$, there exists a continuously differentiable function $\phi: R^{n_1} \rightarrow R^{n_2}$ such that $0 \in R^n$ is globally asymptotically stable with respect to (1.6), and $\det g(x) \neq 0$ for every $x \in R^n$. Then condition (3.10) holds with $W(x) = \frac{1}{2}\|x_2 - \phi(x_1)\|^2$.

Another interesting consequence of Theorem 3.2 is the following result.

THEOREM 3.5. *Assume that there exist a uniformly unbounded and continuously differentiable function $V: R^{n_1} \rightarrow R^+$ for (1.6), a continuous map $\phi: R^{n_1} \rightarrow R^{n_2}$, a strictly increasing continuous function $c: R^+ \rightarrow R^+$ with $c(0) = 0$, and a positive constant K such that*

$$(3.11) \quad DV(x_1)f_1(x_1, \phi(x_1)) \leq -c(\|x_1\|),$$

$$(3.12) \quad \|DV(x_1)\| \leq K\|x_1\|, \quad \forall x_1 \in R^{n_1}.$$

Furthermore, assume that there exist a continuously differentiable function $W: R^n \rightarrow R$ satisfying (2.1), (2.2), and (3.1), and a strictly increasing continuous function $d: R^+ \rightarrow R^+$ such that the following holds:

$$(3.13) \quad \left(\frac{\partial W}{\partial x_2} g \right)(x) = 0, \quad x \notin M_\phi \Rightarrow F(W)(x) < -d(\|x_2 - \phi(x_1)\|).$$

Suppose also that there is a positive constant l such that for any sufficiently large $x \notin M_\phi$ with $((\partial W/\partial x_2)g)(x) = 0$ it holds that

$$(3.14) \quad c(\|x_1\|)d(\|x_2 - \phi(x_1)\|) \geq l\|x_1\|^2\|x_2 - \phi(x_1)\|^2.$$

Finally, assume that the map $\partial f_1/\partial x_2$ exists and there is a constant $M > 0$ satisfying

$$(3.15) \quad \left\| \frac{\partial f_1}{\partial x_2}(x) \right\| < M, \quad \forall x \in R^n.$$

Then system (1.3) satisfies condition (B), and so it admits a global clf.

Proof. We first show that there exists a constant $q_0 > 0$ such that for every nonzero $x \notin M_\phi$ with $((\partial W/\partial x_2)g)(x) = 0$ there is a positive constant $q > 0$ ($q > q_0$, respectively, for sufficiently large x) such that

$$(3.16) \quad qf_1(V)(x) < d(\|x_2 - \phi(x_1)\|).$$

Note that because of (3.11), (3.12), and (3.15) we have

$$\begin{aligned} f_1(V)(x) &\leq DV(x_1)f_1(x_1, \phi(x_1)) + \|DV(x_1)\| \|f_1(x_1, \phi(x_1)) - f_1(x_1, x_2)\| \\ &\leq -c(\|x_1\|) + KM\|x_1\| \|x_2 - \phi(x_1)\|. \end{aligned}$$

Therefore it suffices to show that for any $x \notin M_\phi$ with $((\partial W/\partial x_2)g)(x) = 0$ there is a constant $q > 0$ ($q > q_0$, for sufficiently large x) such that

$$(3.17) \quad q(c(\|x_1\|) - KM\|x_1\| \|x_2 - \phi(x_1)\|) + \frac{1}{2}d(\|x_2 - \phi(x_1)\|) > 0.$$

For $x \neq 0$ such that $x_1 = 0$, the previous inequality holds for any positive q . Consider now the case where $x_1 \neq 0$, $x \notin M_\phi$ with $((\partial W/\partial x_2)g)(x) = 0$. Then (3.17) is equivalent to

$$q\|x_1\|^2 \frac{c(\|x_1\|)}{\|x_1\|^2} - qKM\|x_1\| \|x_2 - \phi(x_1)\| + \frac{1}{2}\|x_2 - \phi(x_1)\|^2 \frac{d(\|x_2 - \phi(x_1)\|)}{\|x_2 - \phi(x_1)\|^2} > 0.$$

The latter is fulfilled, provided that q satisfies the following inequality:

$$(3.18) \quad 2K^{-2}M^{-2} \frac{c(\|x_1\|)}{\|x_1\|^2} \frac{d(\|x_2 - \phi(x_1)\|)}{\|x_2 - \phi(x_1)\|^2} > q > 0.$$

In particular, by (3.14), for each sufficiently large x , condition (3.17) holds, provided that q satisfies (3.18) and, in addition, $q > q_0 \stackrel{\text{def}}{=} K^{-2}M^{-2}l$. Using (3.13) and (3.16), it follows that for any $x \notin M_\phi$ with $((\partial W/\partial x_2)g)(x) = 0$, it holds that

$$F(W)(x) < -d(\|x_2 - \phi(x_1)\|) \leq \min\{-qf_1(V)(x), 0\}$$

for some $q > 0$ ($q > q_0$, respectively, for sufficiently large x). Therefore condition (B) is fulfilled, and the proof is completed. \square

4. Global exponential stabilization. In this section we provide a sufficient condition for the existence of a global clf guaranteeing global exponential stabilizability for (1.3). To prove our main result we need the following lemma.

LEMMA 4.1. *Let Φ be a global clf for (1.1), which, in addition, satisfies the following property. There exists a constant $C > 0$ such that for any nonzero $x \in R^n$ with $G(\Phi)(x) = 0$, it holds that $F(\Phi)(x) \leq -C\Phi(x)$. Then for any positive constant $C' < C$ there exists a feedback law $u = k(x)$, which is smooth for $x \neq 0$, such that*

$$\Phi(x(t, x_0)) \leq \Phi(x_0)e^{-C't}, \quad \forall t \geq 0, \quad x_0 \in R^n,$$

where $x(t, x_0)$ denotes the trajectory of (1.4) of time t starting at x_0 .

The proof of the previous lemma follows by using similar arguments to those given in [3] or [24] and is left to the reader.

The following theorem generalizes Theorem 3.6 of Vidyasagar in [27] concerning the exponential case.

THEOREM 4.2. *Suppose that there exists a map $\phi: R^{n_1} \rightarrow R^{n_2}$, a continuously differentiable function $V: R^{n_1} \rightarrow R^+$, and positive constants $M, D, \theta, c, c_1, c_2$, and K such that (3.11) is fulfilled with $c(s) = cs^2$, $s \geq 0$; conditions (3.12) and (3.15) hold; and, furthermore,*

$$(4.1) \quad c_1\|x_1\|^2 \leq V(x_1) \leq c_2\|x_1\|^2,$$

$$(4.2) \quad \|\phi(x_1)\| \leq D\|x_1\|^\theta, \quad \forall x_1 \in R^{n_1}.$$

Also, assume that there exists a continuously differentiable function $W: R^n \rightarrow R^+$ and positive constants d, d_1 , and d_2 such that

$$(4.3) \quad d_1\|x_2 - \phi(x_1)\|^2 \leq W(x) \leq d_2\|x_2 - \phi(x_1)\|^2, \quad \forall x \in R^n,$$

and (3.13) is fulfilled with $d(s) = ds^2$, $s \geq 0$. Then for any sufficiently small constant $q > 0$, the map

$$(4.4) \quad \Phi(x) = qV(x_1) + W(x)$$

is a global clf. Moreover, system (1.3) is globally exponentially stabilizable.

Proof. Similar to the proof of Theorem 3.5, we get

$$(4.5) \quad f_1(V)(x) \leq -c\|x_1\|^2 + KM\|x_1\|\|x_2 - \phi(x_1)\|$$

and let q be a strictly positive constant satisfying $K^{-2}M^{-2}cd > q$. Then we can easily justify that

$$(4.6) \quad q(-c\|x_1\|^2 + KM\|x_1\|\|x_2 - \phi(x_1)\|) \leq \frac{d}{2}\|x_2 - \phi(x_1)\|^2 - \frac{qc}{2}\|x_1\|^2, \quad \forall x \in R^n,$$

and therefore by (3.13), (4.1), and (4.3)–(4.6), it follows that for any nonzero x with $G(\Phi)(x) = ((\partial W/\partial x_2)g)(x) = 0$, we have

$$\begin{aligned} F(\Phi)(x) &= qf_1(V)(x) + F(W)(x) \\ &\leq \left(\frac{d}{2}\|x_2 - \phi(x_1)\|^2 - \frac{qc}{2}\|x_1\|^2 \right) - d\|x_2 - \phi(x_1)\|^2 \\ (4.7) \quad &= -\frac{qc}{2}\|x_1\|^2 - \frac{d}{2}\|x_2 - \phi(x_1)\|^2 \\ &\leq -\frac{qc}{2c_2}V(x_1) - \frac{d}{2d_2}W(x) \leq -C\Phi(x_0), \end{aligned}$$

where

$$C = \min \left\{ \frac{c}{2c_2}, \frac{d}{2d_2} \right\}.$$

Since Φ is positive definite, the previous inequality asserts that Φ is a global clf. Furthermore, by (4.7) and Lemma 4.1, for any positive $C' < C$ there exists a feedback law $u = k(x)$, which is smooth for $x \neq 0$, such that $0 \in R^n$ is globally exponentially stable with respect to (1.4) and, in addition,

$$(4.8) \quad \Phi(x(t, x_0)) \leq \Phi(x_0) e^{-C't}, \quad \forall t \geq 0.$$

Therefore by (4.1), (4.3), and (4.8) we get

$$qc_1\|x_1(t, x_0)\|^2 + d_1\|x_2(t, x_0) - \phi(x_1(t, x_0))\|^2 \leq \Phi(x_0) e^{-C't}, \quad \forall t \geq 0,$$

and so

$$(4.9a) \quad \|x_1(t, x_0)\| \leq \left(\frac{\Phi(x_0)}{qc_1} \right)^{1/2} \exp \left(-\frac{C'}{2} t \right),$$

$$(4.9b) \quad \|x_2(t, x_0)\| \leq \left(\frac{\Phi(x_0)}{d_1} \right)^{1/2} \exp \left(-\frac{C'}{2} t \right) + \|\phi(x_1(t, x_0))\|.$$

The latter, in conjunction with (4.9a) and assumption (4.2), gives

$$(4.10) \quad \|x_2(t, x_0)\| \leq \left(\frac{\Phi(x_0)}{d_1} \right)^{1/2} \exp \left(-\frac{C'}{2} t \right) + D \left(\frac{\Phi(x_0)}{qc_1} \right)^{\theta/2} \exp \left(-\frac{\theta C'}{2} t \right)$$

for any $t \geq 0$ and $x_0 \in R^n$. Hence, by (4.9b) and (4.10), it follows that the origin $0 \in R^n$ is globally exponentially stable with respect to (1.3). \square

Remark 4.3. Note that Theorem 4.2 is a special case of Theorem 3.5. In particular, the assumptions of Theorem 4.2 imply (3.14) with $l = cd$. Note also that conditions (3.11) with $c(s) = cs^2$, $s \geq 0$, (3.12), as well as (4.1), are fulfilled, if we assume that $0 \in R^n$ is globally exponentially stable with respect to (1.6), the map $x_1 \rightarrow f(x_1, \phi(x_1))$ is continuously differentiable, and its derivative is uniformly bounded on R^n . (See, for instance, [10].)

Finally, we provide a sufficient condition for the existence of the function W satisfying (3.13) with $d(s) = ds^2$, $s \geq 0$, and inequality (4.3).

PROPOSITION 4.4. Suppose that the mappings F , G , and ϕ are continuously differentiable and there exist continuously differentiable mappings $p_0: R^n \rightarrow R^+$ and $p: R^n \rightarrow R^l$ such that p_0 is nonnegative definite and is uniformly bounded on R^n and, furthermore, the set $M_\phi = \{x \in R^n: x_2 = \phi(x_1)\}$ is globally exponentially stable with respect to the resulting system (2.16). In particular, assume that there exist constants $\alpha > 0$ and $\beta > 0$ such that

$$\|x_2(t, x_0) - \phi(x_1(t, x_0))\| \leq \alpha \exp(-\beta t) \|x_{02} - \phi(x_{01})\|, \quad \forall x_0 \in R^n, \quad t \geq 0,$$

where $x(t, x_0)$ denotes the trajectory of (2.16). Finally, assume that there exists a constant $C > 0$ such that

$$\|(p_0(f_2 - D\phi f_1) + gp)(x)\| < C \|x_2 - \phi(x_1)\|, \quad \forall x \in R^n.$$

Then there exists a continuously differentiable real mapping $W: R^n \rightarrow R^+$ and positive constants d , d_1 , and d_2 satisfying inequalities (3.13) with $d(s) = ds^2$, $s \geq 0$, and (4.3).

Outline of the proof. Using standard argument (see, for instance, [10, p. 274]), it can be easily shown that there exist positive constants T , d_1 , and d_2 such that the map

$$W(x_0) = \int_0^T \|x_2(t, x_0) - \phi(x_1(t, x_0))\|^2 dt, \quad x_0 \in R^n$$

is continuously differentiable on R^n and satisfies (4.3). Furthermore, there is a constant $d > 0$ such that

$$(p_0 F + Gp)(W(x)) \leq -d \|x_2 - \phi(x_1)\|^2, \quad \forall x \in R^n.$$

Then, similar to Theorem 2.4, the latter inequality implies (3.13), and the proof is completed. \square

5. Numerical examples. Next, we illustrate the nature of the theory we developed by three numerical examples. In particular, Examples 5.1 and 5.3 have been carefully devised to provide further insight into Theorems 2.2. and 4.2 and, in addition, to show that our method is applicable to cases not covered by the work of Vidyasagar [27]. Example 5.2 shows the applicability of Theorem 3.2 for global stabilization.

Example 5.1. Consider system (1.3) with $x_1 = w_1 \in R$, $x_2 = (w_2, w_3)' \in R^2$, $f_1(x) = -w_1^3 w_2^2 - (w_1 - w_2^3)^3 + w_3^3$, $f_2(x) = (-\frac{1}{3} w_1^3 + w_3^3, w_3 + w_2^2)'$, and $g(x) = (0, w_2^3)'$. Let $\phi(x_1) = (w_1^{1/3}, 0)'$. Then ϕ is continuous (it fails to be continuously differentiable at zero) and the origin $0 \in R$ is asymptotically stable with respect to $\dot{x}_1 = f_1(x_1, \phi(x_1)) = -x_1^{11/3}$. We define $W(x) = \frac{1}{2}((w_1 - w_2^3)^2 + w_3^2)$ and $M_\phi = \{x \in R^3: x_2 = \phi(x_1)\}$. Then for any $x \notin M_\phi$ with $((\partial W / \partial x_2)g)(x) = w_3^3 = 0$, it follows that $w_3 = 0$, $w_2^3 \neq w_1$, and so $F(W)|_{w_3=0} = -(w_1 - w_2^3)^4 < 0$. Hence the system satisfies (A1) and (A2) and, according to Theorem 2.2, is asymptotically stabilizable by means of a feedback law that is smooth for $x \neq 0$ near zero. Note that the linearization $(A, b) = (DF(0), G(0))$ of the system at zero is completely uncontrollable, equivalently $b = 0$, and the matrix A contains a strictly positive eigenvalue. Hence the system cannot be asymptotically stabilized by means

of a feedback law, which is continuously differentiable near zero. Furthermore, there is not any feedback stabilizer $u = k(x)$ that is continuous (or even bounded) near zero. Indeed, otherwise from the third equation $\dot{w}_3 = w_3(1 + w_3k(x)) + w_2^2$ of the resulting closed-loop system, we can easily justify that there would exist a constant $c > 0$ such that $w_3(t, x_0) \geq c$ for any sufficiently small initial state x_0 with $w_3(0) > 0$ and time t , a contradiction.

Example 5.2. Consider system (1.3) with $x_1 = w_1 \in \mathbb{R}$, $x_2 = (w_2, w_3)' \in \mathbb{R}^2$, $f_1(x) = -w_1^3 + w_1\sigma(w_2, w_3)$, $f_2(x) = (-w_2^3w_1^2 - w_2^5, w_3)'$, and $g = (0, 1)'$. Moreover, assume that σ is Lipschitz continuous and satisfies

$$(5.1) \quad \sigma(0, 0) = 0 \quad \text{and} \quad \lim_{|w_2| \rightarrow +\infty} w_2^4 \frac{1}{\sigma(w_2, 0)} > 0.$$

Let $\phi(x_1) = (0, 0)$, $M_\phi = \{x \in \mathbb{R}^3 : x_2 = \phi(x_1)\}$, $W(x) = \frac{1}{2}(w_2^2 + w_3^2)$, and $V(x_1) = \frac{1}{2}w_1^2$. Then we can easily check that the system satisfies condition (B) and so, according to Theorem 3.2, it is globally asymptotically stabilizable. In particular, for any $x \notin M_\phi$ with $((\partial W / \partial x_2)g)(x) = 0$, we get $w_3 = 0$, $(w_1', w_2') \neq 0$, $F(W)(x) = -w_2^4w_1^2 - w_2^6$, and so by (5.1) there is a constant $q_0 > 0$ such that (3.2) is satisfied for some $q > 0$ ($q > q_0$, for sufficiently large x).

Example 5.3. Finally, consider system (1.3) with $x_1 = w_1 \in \mathbb{R}$, $x_2 = (w_2, w_3)' \in \mathbb{R}^2$, $f_1(x) = -w_1 + w_1^3 - w_2 + w_3$, $f_2(x) = ((w_1^3 - w_2)(1 + 3w_1^2) - 3w_1^3, w_3)'$, and $g(x) = (0, w_3^2)$. Let $\phi(x_1) = (w_1^3, 0)'$, $M_\phi = \{x \in \mathbb{R}^3 : x_2 = \phi(x_1)\}$, $W(x) = \frac{1}{2}((w_2 - w_1^3)^2 + w_3^2)$, and $V(x_1) = \frac{1}{2}x_1^2$. Then zero $0 \in \mathbb{R}$ is globally asymptotically stable with respect to $\dot{x}_1 = f_1(x_1, \phi(x_1)) = -x_1$. In particular,

$$(5.2) \quad DV(x_1)f_1(x_1, \phi(x_1)) = -x_1^2.$$

Furthermore, for any $x \notin M_\phi$ near zero with $((\partial W / \partial x_2)g)(x) = 0$, it follows that $w_3 = 0$, $w_2 \neq w_1^3$, and therefore

$$(5.3) \quad F(W)|_{w_3=0, w_2 \neq w_1^3} = -(w_2 - w_1^3)^2 < 0.$$

Hence the system satisfies (A1) and (A2); therefore it is asymptotically stabilizable. Finally, the functions V , ϕ , and W satisfy conditions (4.1)–(4.3), respectively, and $\partial f_1 / \partial x_2$ is uniformly bounded on \mathbb{R}^3 . Therefore, according to (5.2) and (5.3), the assumptions of Theorem 4.2 are fulfilled with $c(s) = d(s) = s^2$, and so the system is globally exponentially stabilizable. Note that although ϕ is continuously differentiable, there is not any Lipschitz continuous map k satisfying (1.7); therefore Vidyasagar's approach does not work. Furthermore, similar to the case of Example 5.1, it can be shown that the system cannot be asymptotically stabilizable by means of a bounded-near-zero feedback law.

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. TMA, 7 (1983), pp. 1163–1173.
- [4] S. P. BANKS, *Stabilizability of finite- and infinite-dimensional bilinear systems*, IMA J. Math. Control Inform., 3 (1986), pp. 255–276.
- [5] W. M. BOOTHBY AND R. MARINO, *Feedback stabilization of planar nonlinear systems*, Systems Control Lett., 12 (1989), pp. 87–92.

- [6] R. W. BROCKETT, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millman, and H. J. Sussmann, eds., Birkhäuser, Boston, 1983, pp. 181–191.
- [7] C. I. BYRNES AND A. ISIDORI, *New results and counterexamples in nonlinear feedback stabilization*, Systems Control Lett., 12 (1989), pp. 437–441.
- [8] P. E. CROUCH, *Spacecraft attitude control and stabilization: Applications of geometric control theory*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 321–333.
- [9] W. P. DAYAWANSA AND C. F. MARTIN, *Asymptotic stabilization of two dimensional real-analytic systems*, Systems Control Lett., 12 (1989), pp. 205–211.
- [10] W. HAHN, *Stability of Motion*, Springer-Verlag, Berlin, New York, 1967.
- [11] H. HERMES, *On the synthesis of a stabilizing feedback control via Lie algebraic methods*, SIAM J. Control Optim., 18 (1980), pp. 352–361.
- [12] M. KAWSKI, *Stabilization of nonlinear systems in the plane*, Systems Control Lett., 12 (1989), pp. 169–176.
- [13] 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.
- [14] 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.
- [15] R. MARINO, *Feedback stabilization of single-input nonlinear systems*, Systems Control Lett., 10 (1988), pp. 201–206.
- [16] E. D. SONTAG, *Smooth stabilization implies coprime factorization*, IEEE Trans. Automat. Control, 34 (1989), pp. 435–443.
- [17] ———, *A “universal” construction of Artstein's theorem on nonlinear stabilization*, Systems Control Lett., 13 (1989), pp. 117–123.
- [18] ———, *Feedback stabilization of nonlinear systems*, in Proc. of the International Symposium on the Mathematical Theory of Networks and Systems, Vol. 2, 1990, pp. 61–81.
- [19] E. D. SONTAG AND H. J. SUSSMANN, *Remarks on continuous feedback*, in Proc. IEEE Conf. Dec. and Control., Albuquerque, NM, December 1980.
- [20] ———, *Further comments on the stabilizability of the angular velocity of a rigid body*, Systems Control Lett., 12 (1989), pp. 213–217.
- [21] H. J. SUSSMANN, *Subanalytic sets and feedback control*, J. Differential Equations, 31 (1979), pp. 31–52.
- [22] J. TSINIAS, *Sufficient Lyapunovlike conditions for stabilization*, Math. Control Signals Systems, 2 (1989), pp. 343–357.
- [23] ———, *Stabilization of affine in control nonlinear systems*, Nonlinear Anal. TMA, 12 (1988), pp. 1283–1296.
- [24] J. TSINIAS AND N. KALOUPSIDIS, *Output feedback stabilization*, IEEE Trans. Automat. Control, 35 (1990), pp. 951–954.
- [25] J. TSINIAS, *Existence of control Lyapunov functions and applications to state feedback stabilizability of nonlinear systems*, SIAM J. Control Optim., 29 (1991), pp. 457–473.
- [26] ———, *Asymptotic feedback stabilization: A sufficient condition for the existence of control Lyapunov functions*, Systems Control Lett., 15 (1990), pp. 441–448.
- [27] M. VIDYASAGAR, *Decomposition techniques for large-scale systems with nonadditive interactions: Stability and stabilizability*, IEEE Trans. Automat. Control, 25 (1980), pp. 773–779.
- [28] F. W. WILSON, *Smoothing derivatives of functions and applications*, Tech. Report 66–3, Brown University, Providence, RI, 1966.
- [29] ———, *The structure of the level surfaces of a Lyapunov function*, J. Differential Equations, 3 (1967), pp. 323–329.

PARAMETER ESTIMATION FOR DISTRIBUTED EQUATIONS IN PARAMETER-DEPENDENT STATE SPACES: APPLICATIONS TO SHAPE IDENTIFICATION*

PATRICIA K. LAMM†

Abstract. This paper presents an approximation theory for the problem of estimating parameters that appear in distributed systems. The work is a generalization of the ideas of Banks and Ito, [*Control Theory and Advanced Technology*, 46 (1988), pp. 73–90] and is motivated by the need to consider an important class of estimation problems not treated by current theory, specifically, those problems in which the underlying state spaces (and consequently the approximating state spaces) are parameter dependent. Situations of this type frequently arise in applications in which it is of interest to estimate unknown (domain) shapes and boundaries, parameters appearing in boundary conditions, or certain functional coefficients present in “degenerate” partial differential equations.

This paper develops the theoretical ideas behind a parameter-dependent approximation theory, and illustrates the application of these ideas to domain optimization problems.

Key words. parameter estimation, shape identification, parameter-dependent state spaces

AMS(MOS) subject classifications. 35, 65, 49

1. Introduction. The problem of estimating unknown functional parameters that appear in mathematical models of physical processes is one of continuing theoretical interest and wide applicability. In recent years, attention has focused on solution methods for problems governed by distributed systems, motivated partly by the continuing interest in the identification of physical parameters associated with flexible structures (e.g., space structures and antennas), ongoing research in inverse problems associated with oil recovery and exploration, and work involving the development of distributed models for biological systems, to name just a few of the many applications in this area.

The types of unknown parameters sought in an identification procedure typically fall into one of two classes: (1) “coefficient-type” parameters, representing quantities associated with mass, flexibility or stiffness, and various rate parameters; initial conditions and coefficients appearing in simple boundary conditions are also important examples of this class; and (2) “domain-type” parameters, such as the shape of the underlying domain or the boundary of the domain (see [4], [10], [11], [14], [16], [24] and numerous articles in [12], [15], [27]), locations of interfaces lying within the domain [13], [17], [18], [20], or unknown quantities appearing in more complex boundary conditions. The distinction between these two classes of parameters from a mathematical point of view is that, in the latter, the parameters typically become part of the definition of the state space associated with the solution of the model equations; this is less often the case in the former. It is interesting to note, however, that “coefficient-type” parameters *can* lead to a situation in which the construction of parameter-dependent state spaces is unavoidable, as is true, for example, in problems where unknown functional coefficients are known to “degenerate” (for details, see [19], [21]).

* Received by the editors December 5, 1988; accepted for publication (in revised form) April 4, 1991.

† Department of Mathematics, Michigan State University, East Lansing, Michigan 48824-1027. This research was supported in part by the National Science Foundation under grants DMS-8807162 and DMS-8601968, and U. S. Air Force Office of Scientific Research contract AFOSR-ISSA-860079. Parts of the work were completed while the author was a visiting associate professor in the Department of Mathematics, University of Southern California, Los Angeles, California 90089-1113.

Banks and Ito [3] have established a unified theory for the approximation of “coefficient-type” parameters that appear in a wide class of first- and second-order linear distributed systems, a development based on state spaces that do *not* depend on unknown parameters. For the first-order case, they address the general problem of estimating the unknown parameter q in the abstract equation

$$(1.1) \quad \dot{u}(t) = A(q)u(t) + F(t; q), \quad t \in (0, T),$$

$$(1.2) \quad u(0) = u_0(q),$$

defined in a Hilbert space H . The parameter q is assumed to belong to a parameter set \mathcal{Q} , while for each $q \in \mathcal{Q}$, $A(q)$ generates an analytic semigroup on H . To determine the unknown q , it is assumed that observations $\tilde{u}_i \in H$ are available for $u(t_i, q)$, and that the method of determining the parameter is a matter of finding $\bar{q} \in \mathcal{Q}$ that minimizes the least squares fit-to-data criterion

$$(1.3) \quad J(q) = \sum_i |u(t_i, q) - \tilde{u}_i|_H^2.$$

A theoretically sound approximation/estimation theory may be found in [3] for the estimation of an optimal q for this problem. Extensions of this theory to nonlinear systems have subsequently been studied by Banks, Reich, and Rosen in [7], [8]. These papers, along with [3], are noteworthy for their general applicability to a large number of inverse problems that have appeared in the literature, and for the ease with which we may verify the (minimal) assumptions required for each of the theories to be valid. Additionally, the theory allows for the use of a higher-order norm in the definition of J in (1.3) (specifically, $\|\cdot\|_V$, where V is a subspace of H , dense and continuously imbedded in H) and a relaxation of the usual constraints on the space in which parameters are defined (i.e., the space must still be compact, but generally in a weaker topology than is usually stipulated to prove convergence); this is done without imposing further regularity on solutions of the distributed system [3].

In the present work, we consider a further generalization of the framework developed by Banks and Ito to treat those problems in which the underlying state spaces (and very often the approximating state spaces) are parameter dependent. In doing so, we are able to extend the basic theory’s applicability and ease of implementation to a large class of estimation problems where unknown parameters may be of both “coefficient” type (parameters q , in a parameter set \mathcal{Q}) and “boundary” type (parameters p , in a parameter set \mathcal{P}). The estimation problem of interest is to determine the unknown parameter vector $(p, q) \in \mathcal{P} \times \mathcal{Q}$ that appears in the abstract equation

$$(1.4) \quad \dot{u}(t) = A(p, q)u(t) + F(t; p, q), \quad t \in (0, T),$$

$$(1.5) \quad u(0) = u_0(p, q),$$

which is now defined in a parameter-dependent Hilbert space $H(p)$. The corresponding parameter estimation problem, and the associated optimal parameters $(\bar{p}, \bar{q}) \in \mathcal{P} \times \mathcal{Q}$, may be defined using a least squares fit-to-data criterion similar to (1.3). In the subsequent sections, we make precise definitions for this problem and for associated approximating problems, and develop a corresponding convergence theory.

The outline of our exposition is as follows. We begin in §2 by presenting an application of domain shape estimation, which motivates our work. We also generalize this example to the overall problem of estimation, in parameter-dependent state space problems, and indicate the steps that need to be taken to develop our approximation

and convergence theory. In §§3 and 4 we present the definitions and assumptions that form the basis for the theoretical development, and develop separately the theory for first-order and second-order problems. Finally, in §5, we return to the domain shape estimation problem and apply the ideas of the previous sections. Notation throughout is standard. We use the notation $L_2(\Omega)$, $H^1(\Omega)$, $H_0^1(\Omega)$, to designate the standard Sobolev spaces on $\Omega \subset \mathbb{R}^n$ (see, for example, [1]), and $\mathcal{L}(X, Y)$ to denote the space of bounded linear operators defined on X with range in Y .

2. Estimation problems with parameter-dependent state spaces. In this section, we present an application to domain shape estimation, which serves to motivate our work. Additionally, we generalize the basic parameter-independent and parameter-dependent spaces and operators that occur naturally in this example, and indicate the steps needed to develop an estimation/approximation theory.

2.1. Applications to domain shape estimation. We present an example where the unknown parameter is the domain itself, or significant features of the domain, such as interfaces or boundary parameters. Such problems have been considered by numerous authors, for example [4], [17], [18], [20], which are special cases of the example given here. We note that the ideas presented by these authors involve approximation schemes that are equivalent to those presented here, but that the underlying theoretical approaches differ significantly from the framework that is considered in this paper. That is, both in these papers and in the construction taken in the example below, a parameter-dependent coordinate transformation is introduced that facilitates the representation of the unknown (parameter-dependent) domain as the range of a (nonparametrized) *fixed* domain. The difference, however, between our theoretical approach and that taken elsewhere is quickly seen when we look at the steps taken in [4], [17], [18], [20] to verify convergence of the resulting schemes. Specifically, in each of these references, a convergence theory is developed for the *transformed* problem, i.e., for the abstract equation *after* it has been transformed via the change of coordinates to a (much more complicated) equation on the nonparametrized reference domain. Thus, convergence is obtained only after a coupling between the original dynamical equation and the coordinate transformation has been made. For example, in [4] the authors develop a convergence theory for a specific domain estimation problem by directly verifying the assumptions given by Banks and Ito in [3] for the *transformed* equations (because the theory in [3] is only valid on *fixed* domain problems); not only are the assumptions not easily verified, but the calculations involved clearly depend on the particular coordinate transformation selected.

One clear advantage then of the theory presented in this paper is that we are able to *separate* the examination of the basic properties of the *original* abstract evolution equations (1.4), (1.5) from a study of the coordinate transformation and its associated regularity properties. We present here an example that illustrates this point.

We let Ω denote a given (fixed) bounded domain in \mathbb{R}^n and consider the problem of estimating an unknown domain Ω_p , where Ω_p is a subset of Ω . Assuming that observations \tilde{u}_i are given at times t_i in the observation space $H^1(\Omega)$, and that $\chi(p)\tilde{u}_i$ denotes the restriction of \tilde{u}_i to Ω_p , our goal will be to determine a parameter q and a domain Ω_p minimizing

$$(2.1) \quad J_t(p, q) = \sum_i \|u(t_i; p, q) - \chi(p)\tilde{u}_i\|_{H^1(\Omega_p)}^2$$

over all possible q , $q = (a_{i,j}, i, j = 1, \dots, n; a_j, j = 1, \dots, n; a_0) \in \mathcal{Q}$, and all possible choices of Ω_p (i.e., all $p \in \mathcal{P}$, where p is used to parametrize Ω_p); the sets \mathcal{P}

and \mathcal{Q} will be characterized shortly. In (2.1), the state variable $u(\cdot; p, q)$ satisfies the model equations

$$(2.2) \quad \frac{\partial u}{\partial t} = \sum_{i,j=1}^n \partial_j (a_{ij} \partial_i u) + \sum_{j=1}^n a_j \partial_j u + a_0 u + f(t) \quad \text{on } \Omega_p, \quad t \in (0, T),$$

$$(2.3) \quad u(0) = \hat{u}_0 \quad \text{on } \Omega_p,$$

as well as homogeneous boundary conditions of either Dirichlet or Neumann type. Here ∂_i denotes the partial differentiation operator, with respect to the i th spatial variable.

Such an estimation problem occurs in many applications, for example, the many optimal shape design problems for engines, ships, and airplanes given in [24]. To consider a specific example, Banks and Kojima [4] consider the problem of the detection and characterization of large structural flaws (not observable in visual inspections) occurring in aerospace structures. The flaw, when it occurs, is assumed to alter the usual shape of a specific layer within layered composite media. The physical estimation problem involves applying a heat source to the material, and collecting heat measurements that are the result of heat conduction through the layered media; these measurements are then used to determine the true values of thermal diffusivity parameters for the material. The definition of the diffusivity parameters, and of the corresponding parabolic equations for heat conduction, depend on the (unknown) shape of the layer in question. Thus, through measurements of heat on the boundary of a larger, fixed domain (the domain determined by the overall shape of the object being tested), we desire to estimate the shape of an internal region that is not directly available to visual inspection. The problem we consider here is of the form considered in [4]; although we do not directly address the use of boundary measurements, there is no difficulty in using such observations, provided the domain of interest is modeled in \mathbb{R}^1 or \mathbb{R}^2 .

We next prescribe conditions on the parameter sets \mathcal{P} and \mathcal{Q} . For now, we will assume that the parameter set \mathcal{Q} for q is such that all coefficients are given in $L_\infty(\Omega)$ and that, for some $c > 0$, each $q \in \mathcal{Q}$ satisfies

$$\operatorname{Re} \sum_{i,j=1}^n a_{ij}(x) \xi_i \bar{\xi}_j \geq c \sum_{j=1}^n |\xi_j|^2, \quad \xi = (\xi_1, \dots, \xi_n) \in \mathcal{C}^n, \quad x \in \Omega.$$

In general, for domain shape estimation problems, the parameter set \mathcal{P} consists of parameters p appearing in some a priori parameterization of the curved boundary of Ω_p , a representation dependent on the particular geometry and application being considered; it will not be important here to know precisely how Ω_p is defined using p ; rather, it suffices to assume that the relationship between Ω_p and $p \in \mathcal{P}$ is such that the following two conditions are satisfied:

1. For each $p \in \mathcal{P}$, the boundary of Ω_p satisfies a strong local Lipschitz condition [1].
2. The topology on \mathcal{P} is defined so that the following property holds: for any $\epsilon > 0$, there exists $\delta > 0$ such that, given arbitrary $p, \tilde{p} \in \mathcal{P}$ satisfying $d_p(\tilde{p}, p) < \delta$, we have

$$\int_{\mathcal{R}(\Omega_p, \Omega_{\tilde{p}})} d\Omega < \epsilon,$$

where $\mathcal{R}(\Omega_p, \Omega_{\bar{p}}) \equiv (\Omega_p \cup \Omega_{\bar{p}}) \setminus (\Omega_p \cap \Omega_{\bar{p}})$ and d_p is the metric on \mathcal{P} . Roughly speaking, then, if $d_p(\bar{p}, p) \approx 0$, the regions Ω_p and $\Omega_{\bar{p}}$ “nearly overlap.”

Finally, for reasons that will be given in later sections, we assume that \mathcal{P} and \mathcal{Q} are compact in their corresponding topologies.

The goal then is to determine an “optimal” (\bar{p}, \bar{q}) minimizing J_t in (2.1) over $\mathcal{P} \times \mathcal{Q}$, where $u(\cdot; p, q)$ in (2.1) satisfies (2.2), (2.3). Our general approach will be to define a cost functional J_t^N approximating J_t , where J_t^N is associated with finite-dimensional equations approximating (2.2), (2.3); it is then possible to minimize J_t^N over an “approximate” space $\mathcal{P}^m \times \mathcal{Q}^m$ for a solution $(\bar{p}^{N,m}, \bar{q}^{N,m})$ that approximates (\bar{p}, \bar{q}) in some sense.

2.2. Generalization of the shape estimation problem. Before turning to a theory for parameter-dependent state spaces, it is worthwhile to first restate the domain shape estimation problem in terms of general operators and spaces.

First, we are given parameter sets \mathcal{P}, \mathcal{Q} , where, in general, we assume that $\mathcal{P} \subset \tilde{\mathcal{P}}, \mathcal{Q} \subset \tilde{\mathcal{Q}}$, where $\tilde{\mathcal{P}}$ and $\tilde{\mathcal{Q}}$ are metric spaces, and that \mathcal{P}, \mathcal{Q} are compact in their respective topologies.

Second, for the domain estimation problem we made the assumption that the unknown domain Ω_p was known to be contained in a *fixed* region Ω (given a priori); such an assumption immediately defines both parameter-dependent and parameter-independent spaces. The need for both types of spaces is dictated by the fact that observations \tilde{u}_i are given on the larger domain, while state variables satisfying (2.2), (2.3) are defined on the smaller p -dependent domain. The parameter-independent state and observation spaces, and corresponding parameter-dependent state/observation spaces are given as follows:

- We define the Hilbert spaces $(H, \langle \cdot, \cdot \rangle, | \cdot |)$ and $(V, (\cdot, \cdot), \| \cdot \|)$, V a subspace of H , where H denotes a (fixed) state space and V denotes a (fixed, typically more regular) observation space. For the application considered here, $H \equiv L_2(\Omega)$ and $V \equiv H^1(\Omega)$.
- We define parameter-dependent state spaces $(H(p), \langle \cdot, \cdot \rangle_p, | \cdot |_p)$, and observation spaces, $(V(p), (\cdot, \cdot)_p, \| \cdot \|_p)$, where $H(p), V(p)$ are Hilbert spaces, $V(p) \subset H(p)$; for our example, $H(p)$ and $V(p)$ are the p -dependent analogues of H and V , $H(p) \equiv L_2(\Omega_p)$, and $V(p) \equiv H^1(\Omega_p)$.

Because the least squares fit-to-data criterion J_t makes a comparison between model state variables with domain Ω_p and observations given on Ω (via restriction of observations to the smaller domain), a mapping between parameter-independent and parameter-dependent spaces is naturally given in this matching. That is, we may define the “data truncation map” $\pi_t(p)$, where

$$\pi_t(p) : H \rightarrow H(p), \quad \pi_t(p) : V \rightarrow V(p),$$

and rewrite the fit-to-data criterion (2.1) as

$$(2.4) \quad J_t(p, q) = \sum_i \|u(t_i; p, q) - \pi_t(p)\tilde{u}_i\|_p^2.$$

We note that the truncation map $\pi_t(p)$ is defined according to the particular application; i.e., this particular map defines how observations that are given on V are to be properly viewed as elements in $V(p)$ for each $p \in \mathcal{P}$. For the domain estimation problem considered here, $\pi_t(p) = \chi(p)$.

Remark 2.1. Corresponding to the “truncation” map, we may often define its corresponding right inverse, or a natural “extension” map. This is easily done for the

domain estimation problem; indeed, the fact that Ω_p satisfies a strong local Lipschitz property guarantees the existence of a map $\pi_e(p) \in \mathcal{L}(H(p), H) \cap \mathcal{L}(V(p), V)$ with the property that, for any $u \in H(p) = L_2(\Omega_p)$, $\pi_e(p)u(x) = u(x)$, almost all x in Ω_p [1]. Using π_e instead of π_t , we could have also defined, in contrast to the original least squares functional J_t in (2.4), an alternate fit-to-data criterion J_e

$$J_e(p, q) = \sum_i \|\pi_e(p)u(t_i; p, q) - \tilde{u}_i\|^2,$$

depending on whether it makes more sense (in the context of the application) to “truncate” data or to “extend” solutions to accomplish the matching. For the purposes of the domain estimation example, we will only consider J_t .

Regardless of whether the original J_t or the alternate J_e is used to define the parameter estimation problem, the notion of “data truncation” or “data extension” maps will not be needed in the theoretical development to follow in later sections; rather, we use the notion of a combined “data space-changing map,”

$$\pi(\tilde{p}, p) : H(p) \rightarrow H(\tilde{p}), \quad \pi(\tilde{p}, p) : V(p) \rightarrow V(\tilde{p}),$$

where, clearly, for the example at hand, $\pi(\tilde{p}, p) = \pi_t(\tilde{p})\pi_e(p)$, $\pi(\tilde{p}, p) \in \mathcal{L}(H(p), H(\tilde{p})) \cap \mathcal{L}(V(p), V(\tilde{p}))$.

Writing $F(\cdot; p) = \pi_t(p)f(\cdot)$ and $u_0(p) = \pi_t(p)\hat{u}_0$, the parabolic equations (2.2), (2.3) for the shape estimation problem may be written as a special case of the abstract equations

$$(2.5) \quad \dot{u}(t) = A(p, q)u(t) + F(t; p, q), \quad t \in (0, T),$$

$$(2.6) \quad u(0) = u_0(p, q),$$

in the Hilbert space $H(p)$. The linear operator $A(p, q)$ appearing in (2.5), (2.6) is given uniquely by the sesquilinear form $\sigma(p, q)$, where

$$(2.7) \quad \sigma(p, q)(u, v) \equiv \int_{\Omega_p} \left\{ \sum_{i,j=1}^n a_{ij} \partial_i u \partial_j \bar{v} + \sum_{j=0}^n a_j \partial_j u \bar{v} + a_0 u \bar{v} \right\} d\Omega,$$

for all $u, v \in V_0(p)$, where $V_0(p) \equiv H_0^1(\Omega_p)$ for Dirichlet boundary conditions and $V_0(p) \equiv H^1(\Omega_p)$ for Neumann boundary conditions; that is, for the general case we are given, for all $(p, q) \in \mathcal{P} \times \mathcal{Q}$, a Hilbert space $V_0(p)$, $V_0(p)$ a closed subspace of $V(p)$, as well as a sesquilinear form

$$\sigma(p, q) : V_0(p) \times V_0(p) \rightarrow \mathcal{C},$$

and a corresponding operator $A(p, q)$, where

$$\text{dom } A(p, q) = \{u_0 \in V_0(p) \mid |\sigma(p, q)(u_0, v_0)| \leq c_u |v_0|_p, \text{ all } v_0 \in V_0(p)\}$$

and $A(p, q)$ satisfies

$$\sigma(p, q)(u_0, v_0) = \langle -A(p, q)u_0, v_0 \rangle_p$$

for all $u_0 \in \text{dom } A(p, q)$ and $v_0 \in V_0(p)$. In addition, corresponding to the new space $V_0(p)$ in $V(p)$, we define the parameter-independent analogue V_0 , $V_0 \subset V$ Hilbert; for

the domain estimation problem, V_0 is either $H_0^1(\Omega)$ or $H^1(\Omega)$ (depending on boundary conditions in $V_0(p)$).

Estimation problems associated with distributed system models, such as the domain shape estimation problem, typically have parameters (p, q) and the state variable $u(\cdot; p, q)$ belonging to function spaces so that the estimation problem considered is generally infinite-dimensional in nature. However, if we are able to construct a discretized approximation for $u(t_i; p, q)$, the minimization problem then becomes a finite-dimensional one (assuming for the moment that \mathcal{P} and \mathcal{Q} are finite dimensional; see Remark 2.2). The approximation task for the state variable $u(t; p, q)$ will be accomplished here and in the sections that follow through the definition of Galerkin approximations $u^N(t; p, q)$ in finite-dimensional subspaces $H^N(p)$ of $H(p)$; that is, for each $p \in \mathcal{P}$, $H^N(p) \subseteq V_0(p)$, and $u^N(\cdot; p, q)$ satisfies

$$(2.8) \quad \dot{u}^N(t) = A^N(p, q)u^N(t) + P^N(p)F(t; p, q), \quad t \in (0, T),$$

$$(2.9) \quad u^N(0) = P^N(p)u_0(p, q).$$

Here $A^N(p, q)$ is the operator defined by the restriction of $\sigma(p, q)$ to $H^N(p) \times H^N(p)$, and $P^N(p) : H \rightarrow H^N(p)$ is the orthogonal projection. The finite-dimensional estimation problem then becomes that of determining (\bar{p}^N, \bar{q}^N) , which minimizes one of the two approximate cost functionals, $J_t^N(p, q)$ or $J_e^N(p, q)$, over $\mathcal{P} \times \mathcal{Q}$,

$$(2.10) \quad J_t^N(p, q) = \sum_i \|u^N(t_i; p, q) - \pi_t(p)\tilde{u}_i\|_p^2,$$

$$(2.11) \quad J_e^N(p, q) = \sum_i \|\pi_e(p)u^N(t_i; p, q) - \tilde{u}_i\|^2.$$

The construction of approximating equations (2.8), (2.9), and approximating least squares functionals (2.10) or (2.11) is standard, once the approximating spaces $H^N(p)$ have been defined. What is important to note is that, for many examples of the type considered here, it is desirable to build into the approximating spaces a natural dependence on the unknown parameter p . For example, approximation spaces that lead to particularly efficient computation schemes (see [4], [17], [18], [20] for a discussion of implementation) are given via the following:

- Define a family H^N of *fixed* finite-dimensional approximation spaces, $H^N \subset V_0$, associated with the *fixed* state space H ; for example, a family of spline-based approximations may be used, so long as the order of splines is such that elements are of sufficient smoothness and satisfy needed boundary conditions. (See, for example, [3], [4], [6] and §5. Complete conditions on approximation spaces are given in the next section.)
- Define finite-dimensional parameter-dependent spaces $H^N(p)$ by $H^N(p) \equiv \gamma_t(p)H^N$, where $\gamma_t(p)$ is a smooth mapping (typically, a coordinate transformation) from V_0 to $V_0(p)$.

We consider the definition of $\gamma_t(p)$ for a general parameter-dependent state-space problem. Although such a map will typically be an isomorphism between parameter-independent and parameter-dependent spaces, $\gamma_t(p)$ need not be invertible and is defined solely to implement approximation and to effectively move from one space to another. We note that these maps (to be henceforth called “theoretical space-changing maps”) differ from the previously defined “data space-changing maps” both in definition and in use.

For the domain shape estimation example, the “theoretical” space-changing maps $\gamma_t(p)$ and $\gamma_e(p)$ are defined by first prescribing a one-to-one onto coordinate transformation $\phi(p) : \Omega \rightarrow \Omega_p$, whose inverse will be denoted by $\psi(p) \equiv (\phi(p))^{-1}$. This transformation is assumed to be 1-smooth [1] (i.e., for $i = 1, \dots, n$, $y_i \equiv \phi_i(p)(x)$ and $x_i \equiv \psi_i(p)(y)$ satisfy $\phi_i \in C^1(\bar{\Omega})$, $\psi_i \in C^1(\bar{\Omega}_p)$, $\det(\phi(p)'(x)) \geq k_1 > 0$, $p \in \mathcal{P}$, $x \in \Omega$) and, in the case of Dirichlet boundary conditions, we further require that $\phi(p) : \partial\Omega \rightarrow \partial\Omega_p$ as an onto map. Not surprisingly, it will later be important that we also require that the coordinate transformations vary continuously with p and thus assume that for arbitrary $\epsilon > 0$, there exists $\delta = \delta(\epsilon)$ such that for any $p, \tilde{p} \in \mathcal{P}$ with $d_p(\tilde{p}, p) < \delta$ we have

$$|\phi(p) - \phi(\tilde{p})|_{W_\infty^1(\Omega)} < \epsilon$$

for $i = 1, \dots, n$. (We note that this assumption, coupled with the compactness of \mathcal{P} , guarantees uniform boundedness of the Jacobian for the transformation; i.e., there exist positive constants k_1, K_1 , such that for all $p \in \mathcal{P}$ we have $k_1 \leq \det(\phi(p))'(x) \leq K_1$, almost all $x \in \bar{\Omega}$.) Using the coordinate transformation map, we then define for the domain shape estimation example a “theoretical truncation map” $\gamma_t(p)$ by

$$\gamma_t(p)u \equiv u \circ \psi(p), \quad u \in L_2(\Omega)$$

and a “theoretical extension map” by $\gamma_e(p) \equiv \gamma_t(p)^{-1}$; approximating spaces are then defined for this example from H^N using $\gamma_t(p)$, as already indicated.

Under the continuous dependence with respect to parameters of $u_0(p, q)$, $F(\cdot; p, q)$, the steps by which we may argue (a) the existence of a minimizer (\bar{p}^N, \bar{q}^N) for J_t^N , (J_e^N) , (b) the existence of a “true” (optimal) parameter (\bar{p}, \bar{q}) , and (c) the sense in which (\bar{p}^N, \bar{q}^N) approximates (\bar{p}, \bar{q}) , are by now standard (see, for example, [2]). An approximation/convergence theory is obtained once we illustrate the continuous dependence of the maps $(p, q) \rightarrow J_t^M(p, q)$ (or $(p, q) \rightarrow J_e^M(p, q)$) for every M , and the convergence $J_t^N(p^N, q^N) \rightarrow J_t(p, q)$ (or $J_e^N(p^N, q^N) \rightarrow J_e(p, q)$) whenever $(p^N, q^N) \rightarrow (p, q)$. Using the ideas indicated in Remark 2.3 below, these two results may be established by taking the following steps:

1. *Establish the continuous dependence, for each t , of $u^M(t; p, q) \in H^M(p)$ on (p, q) :* For each fixed $M = 1, 2, \dots$ and arbitrary $(p^N, q^N) \rightarrow (p, q) \in \mathcal{P} \times \mathcal{Q}$,

$$(2.12) \quad \|\pi(p^N, p)u^M(t; p, q) - u^M(t; p^N, q^N)\|_{p^N} \rightarrow 0, \quad \text{as } N \rightarrow \infty.$$

2. *Establish state variable convergence:* For arbitrary $(p^N, q^N) \rightarrow (p, q) \in \mathcal{P} \times \mathcal{Q}$,

$$(2.13) \quad \|\pi(p^N, p)u(t; p, q) - u^N(t; p^N, q^N)\|_{p^N} \rightarrow 0, \quad \text{as } N \rightarrow \infty.$$

In the sections that follow, we will state the assumptions required to verify steps 1 and 2 above for parameter estimation problems when state spaces are parameter dependent. There we will clearly define what is meant by approximating equations and approximate estimation problems; furthermore, conditions will be given on spaces H , V , V_0 , $H(p)$, $V(p)$, $V_0(p)$, $H^N(p)$, and maps $\sigma(p, q)$, $\pi(p, \tilde{p})$, $\gamma_t(p)$, and $\gamma_e(p)$, which ensure that a combined approximation/estimation theory holds. Not surprisingly, many of these conditions will involve a type of “continuous dependence with respect to parameters” of these spaces and operators.

Remark 2.2. Unless the parameter sets \mathcal{P} and \mathcal{Q} are already discrete, the minimization of J_t^N (J_e^N) over $\mathcal{P} \times \mathcal{Q}$ remains an infinite-dimensional problem. It is not

difficult to define a fully discrete problem for this situation, as numerous authors have done (see for example, [6]); we summarize here the basic requirements of such a construction using the steps outlined in [7]. For each $m = 1, 2, \dots$, we define maps $I_{\mathcal{P}}^m : \mathcal{P} \subseteq \tilde{\mathcal{P}} \rightarrow \tilde{\mathcal{P}}$, $I_{\mathcal{Q}}^m : \mathcal{Q} \subseteq \tilde{\mathcal{Q}} \rightarrow \tilde{\mathcal{Q}}$, which satisfy the conditions (i) $I_{\mathcal{P}}^m$, $I_{\mathcal{Q}}^m$ are continuous, (ii) the ranges of $I_{\mathcal{P}}^m$, $I_{\mathcal{Q}}^m$ are finite-dimensional, and (iii) $I_{\mathcal{P}}^m(p) \rightarrow p$, $I_{\mathcal{Q}}^m(q) \rightarrow q$ as $m \rightarrow \infty$ at rates uniform in $(p, q) \in \mathcal{P} \times \mathcal{Q}$. It follows from the compactness of \mathcal{P} and \mathcal{Q} that $\mathcal{P}^m \times \mathcal{Q}^m \equiv I_{\mathcal{P}}^m(\mathcal{P}) \times I_{\mathcal{Q}}^m(\mathcal{Q})$ is compact, so that (using the results in step 1 above) a solution $(\bar{p}^{N,m}, \bar{q}^{N,m})$ exists for the problem of minimizing $J_t^N(J_e^N)$ over $\mathcal{P}^m \times \mathcal{Q}^m$; we note that this minimization is now a fully discrete problem for which standard iterative search techniques may be used. Furthermore, the convergence established in step 2 may be used to argue (subsequential) convergence of the iterates $(\bar{p}^{N,m}, \bar{q}^{N,m})$ to an optimal parameter (\bar{p}, \bar{q}) minimizing $J_t(J_e)$ over $\mathcal{P} \times \mathcal{Q}$.

Remark 2.3. We briefly illustrate here how the verification of step 2 above leads to a statement of the convergence $J_t^N(p^N, q^N) \rightarrow J_t(p, q)$ ($J_e^N(p^N, q^N) \rightarrow J_e(p, q)$) whenever $(p^N, q^N) \rightarrow (p, q)$. (Similar arguments may be made to demonstrate how step 1 above leads to a statement of continuous dependence of $J_t^M(p, q)$, or $J_e^M(p, q)$, on (p, q) .) Additional assumptions are required, depending on whether J_t or J_e is used; we indicate each assumption here, as it is introduced.

For a theory using J_t as the data-fitting functional, we assume that $\pi_t(p)\pi_e(p)$ is the identity map and apply the triangle inequality to obtain

$$\begin{aligned} & |J_t(p, q) - J_t^N(p^N, q^N)| \\ & \leq \sum_i \left| \|\pi_t(p)(\pi_e(p)u(t_i; p, q) - \tilde{u}_i)\|_p - \|\pi_t(p^N)(\pi_e(p)u(t_i; p, q) - \tilde{u}_i)\|_{p^N} \right| \\ & \quad + \sum_i \left| \|\pi(p^N, p)u(t_i; p, q) - \pi_t(p^N)\tilde{u}_i\|_{p^N} - \|u^N(t_i; p^N, q^N) - \pi_t(p^N)\tilde{u}_i\|_{p^N} \right|. \end{aligned}$$

An additional assumption must be made here, namely that of continuity of the map $p \rightarrow \|\pi_t(p)v\|_p : \mathcal{P} \rightarrow \mathbb{R}$ for arbitrary $v \in V$. This is easily seen to hold for the example of domain shape estimation problems (see §5). Using this assumption, it follows that the first term after the above inequality converges to zero as $N \rightarrow \infty$, while the second term is bounded above by $\sum_i \|\pi(p^N, p)u(t_i; p, q) - u^N(t_i; p^N, q^N)\|_{p^N}$. It is the convergence of this last expression that is the subject of step 2, above.

If, instead, the “extension” form J_e of the cost functional is used, we observe that

$$\begin{aligned} |J_e(p, q) - J_e^N(p^N, q^N)| & \leq \sum_i \|\pi_e(p)u(t_i; p, q) - \pi_e(p^N)u^N(t_i; p^N, q^N)\| \\ & \leq \sum_i \|(\pi_e(p) - \pi_e(p^N)\pi_t(p^N)\pi_e(p))u(t_i; p, q)\| \\ & \quad + \sum_i \|\pi_e(p^N)(\pi_t(p^N)\pi_e(p)u(t_i; p, q) - u^N(t_i; p^N, q^N))\|. \end{aligned}$$

Under the (reasonable) assumption that the first term in the last expression converges to zero as $N \rightarrow \infty$, and the condition that $\pi_e(p^N)$ is uniformly bounded (in the operator norm) for all N , we again see that the convergence $J_e^N(p^N, q^N) \rightarrow J_e(p, q)$ is assured if step 2 above has been verified.

3. An approximation framework for first order problems. Let V_0 , V , and H be (nonparameter-dependent) Hilbert spaces with $V_0 \subseteq V \subseteq H$, where it is assumed

that V_0 is dense and continuously imbedded in H , and V_0 is a closed (not necessarily proper) subspace of V . As is discussed in the previous section, we let \mathcal{P} and \mathcal{Q} be compact in given metric spaces $(\tilde{\mathcal{P}}, d_p)$ and $(\tilde{\mathcal{Q}}, d_q)$, respectively, and for each $p \in \mathcal{P}$, let $V_0(p)$, $V(p)$, and $H(p)$ denote Hilbert spaces where the containment

$$V_0(p) \subseteq V(p) \subseteq H(p)$$

is exactly as it is for the nonparametrized spaces, and (as for those spaces) $V_0(p)$ is dense and continuously imbedded in $H(p)$ and forms a closed subspace of $V(p)$.

We let $|\cdot|$ and $\|\cdot\|$ represent the norms for H and V , respectively, while $|\cdot|_p$, $\|\cdot\|_p$ ($\langle \cdot, \cdot \rangle_p$, $(\cdot, \cdot)_p$) will be used to denote the norms (inner products) for $H(p)$ and $V(p)$, respectively. In addition, we use the designation $\mathcal{L}(X, Y)$ to represent the space of bounded linear operators that are defined on the normed linear space X and have range in Y .

We make the following standing hypotheses.

Hypothesis 1 (Assumptions about spaces).

(a) *Uniform (in p) norm constants*: There exists $k > 0$ such that for every $p \in \mathcal{P}$ and all $v \in V(p)$, $|v|_p \leq k\|v\|_p$.

(b) *“Theoretical” space-changing map γ* : For each $p, \tilde{p} \in \mathcal{P}$, we define $\gamma(\tilde{p}, p) \equiv \gamma_t(\tilde{p})\gamma_e(p)$, where we assume that γ_t and γ_e satisfy the following conditions.

(i) *(Theoretical) “Truncation” map γ_t* : The map $\gamma_t(p) \in \mathcal{L}(H, H(p))$ is such that $\gamma_t(p)$ restricted to V belongs to $\mathcal{L}(V, V(p))$, $\gamma_t(p) : V_0 \rightarrow V_0(p)$. In addition, there exists $K \geq 0$ such that for any $p \in \mathcal{P}$,

$$\begin{aligned} \|\gamma_t(p)v\|_p &\leq K\|v\|, & v \in V, \\ |\gamma_t(p)z|_p &\leq K|z|, & z \in H. \end{aligned}$$

(ii) *(Theoretical) “extension” map γ_e* : The map $\gamma_e(p) \in \mathcal{L}(H(p), H)$ is such that $\gamma_e(p)$ restricted to $V(p)$ belongs to $\mathcal{L}(V(p), V)$. Furthermore, there exists $K \geq 0$ such that for any $p \in \mathcal{P}$,

$$\begin{aligned} \|\gamma_e(p)v\| &\leq K\|v\|_p, & v \in V(p), \\ |\gamma_e(p)z| &\leq K|z|_p, & z \in H(p). \end{aligned}$$

(iii) *Continuity properties of theoretical maps γ_t and γ_e* : There exists $C_\gamma : \mathcal{P} \times \mathcal{P} \rightarrow \mathbb{R}^+$ with the property that $C_\gamma(p, \tilde{p}) < \epsilon$ whenever $d_p(p, \tilde{p}) < \delta = \delta(\epsilon)$, and, for any $p, \tilde{p} \in \mathcal{P}$ and arbitrary $v \in V$,

$$(3.1) \quad \|(\gamma_t(\tilde{p}) - \gamma_t(\tilde{p})\gamma_e(p)\gamma_t(p))v\|_{\tilde{p}} \leq C_\gamma(p, \tilde{p})\|v\|.$$

(iv) *Composite map preserves V_0 spaces*: For every $p, \tilde{p} \in \mathcal{P}$, $\gamma(\tilde{p}, p) : V_0(p) \rightarrow V_0(\tilde{p})$.

(v) *Composite map as identity*: For every $p \in \mathcal{P}$, $\gamma(p, p) = I$, the identity on $H(p)$.

(c) *“Data” space-changing map π* : For each $p, \tilde{p} \in \mathcal{P}$, there exists $\pi(\tilde{p}, p) \in \mathcal{L}(H(p), H(\tilde{p}))$ such that $\pi(\tilde{p}, p)$ restricted to $V(p)$ is in $\mathcal{L}(V(p), V(\tilde{p}))$. In addition, given $\epsilon > 0$, $p \in \mathcal{P}$, and $v_0 \in V_0(p)$, there exists $\delta = \delta(\epsilon, p, v_0) > 0$ such that for any $\tilde{p} \in \mathcal{P}$ satisfying $d_p(\tilde{p}, p) < \delta$, we have

$$\|(\pi(\tilde{p}, p) - \gamma(\tilde{p}, p))v_0\|_{\tilde{p}} < \epsilon.$$

(d) *Continuity (in p) of $H(p)$ inner product*: For each $p \in \mathcal{P}$, there exists a map $C_p : \mathcal{P} \rightarrow \mathbb{R}^+$, which is continuous at p and is defined such that for any $\tilde{p} \in \mathcal{P}$, we have

$$\left| \langle \gamma_t(\tilde{p})x, \gamma_t(\tilde{p})z \rangle_{\tilde{p}} - \langle \gamma_t(p)x, \gamma_t(p)z \rangle_p \right| \leq C_p(\tilde{p})|x||z|,$$

for any $x, z \in H$.

Remark 3.1. The condition of uniform norm constants in (a) of Hypothesis 1 occurs automatically if the spaces $H(p)$ and $V(p)$ are defined as the range under $\gamma_t(p)$ of H and V , respectively, provided that $\gamma_t(p)$ is an isomorphism with inverse uniformly bounded in the $\mathcal{L}(V(p), V)$ operator norm.

Remark 3.2. If $V_0(\tilde{p}) \neq V(\tilde{p})$, then $\pi(\tilde{p}, p)v_0$ need only belong to $V(\tilde{p})$, and *not* to $V_0(\tilde{p})$. Thus, (c) of Hypothesis 1 guarantees that, if we *start* from $v_0 \in V_0(p)$, then for \tilde{p} close to p , the distance from $\pi(\tilde{p}, p)v_0$ to $V_0(\tilde{p})$ will be small. Using density of $V_0(p)$ in $H(p)$, we may also obtain from this a similar result for $|(\pi(\tilde{p}, p) - \gamma(\tilde{p}, p))z|_{\tilde{p}}$, given arbitrary z in $H(p)$.

For $(p, q) \in \mathcal{P} \times \mathcal{Q}$ given, let $\sigma(p, q) : V_0(p) \times V_0(p) \rightarrow \mathcal{C}$ denote a sesquilinear form that satisfies the following assumptions.

Hypothesis 2 (Assumptions about sesquilinear form $\sigma(p, q)$).

(a) *Continuity in parameters.*

(i) *Continuity in p* : For any $q \in \mathcal{Q}$ and $p \in \mathcal{P}$, there exists a map $C_{p,q} : \mathcal{P} \rightarrow \mathbb{R}^+$ that is continuous at p and is defined such that for every $\tilde{p} \in \mathcal{P}$,

$$\begin{aligned} & \left| \sigma(\tilde{p}, q)(\gamma_t(\tilde{p})u_0, \gamma_t(\tilde{p})v_0) - \sigma(p, q)(\gamma_t(p)u_0, \gamma_t(p)v_0) \right| \\ & \leq C_{p,q}(\tilde{p})\|u_0\| \|v_0\|, \end{aligned}$$

for every $u_0, v_0 \in V_0$.

(ii) *Uniform (in p) q -continuity*: For every $q \in \mathcal{Q}$, there exists a map $C_q : \mathcal{Q} \rightarrow \mathbb{R}^+$ that is continuous at q and is defined such that for every $p \in \mathcal{P}$, $\tilde{q} \in \mathcal{Q}$,

$$|\sigma(p, q)(u_0, v_0) - \sigma(p, \tilde{q})(u_0, v_0)| \leq C_q(\tilde{q})\|u_0\|_p \|v_0\|_p,$$

for every $u_0, v_0 \in V_0(p)$.

(b) *Uniform (in (p, q)) $V_0(p)$ -boundedness*: There exists $c_2 > 0$ such that for every $(p, q) \in \mathcal{P} \times \mathcal{Q}$,

$$|\sigma(p, q)(u_0, v_0)| \leq c_2\|u_0\|_p \|v_0\|_p,$$

for every $u_0, v_0 \in V_0(p)$.

(c) *Uniform (in (p, q)) $V_0(p)$ -coercivity*: There exists $c_1 > 0$ and some $\lambda_0 \in \mathbb{R}$ such that for every $(p, q) \in \mathcal{P} \times \mathcal{Q}$ and all $v_0 \in V_0(p)$,

$$\operatorname{Re} \sigma(p, q)(v_0, v_0) + \lambda_0|v_0|_p^2 \geq c_1\|v_0\|_p^2.$$

It is well known [25], [23] that the conditions of $V_0(p)$ -coercivity and boundedness of $\sigma(p, q)$ are sufficient to guarantee that $\sigma(p, q)$ defines a linear operator $A(p, q)$ with $\operatorname{dom} A(p, q) = \{u_0 \in V_0(p) \mid |\sigma(p, q)(u_0, v_0)| \leq c_u|v_0|_p, \text{ all } v_0 \in V_0(p)\}$, such that $A(p, q)$ satisfies

$$\sigma(p, q)(u_0, v_0) = \langle -A(p, q)u_0, v_0 \rangle_p$$

for all $u_0 \in \text{dom } A(p, q)$ and $v_0 \in V_0(p)$. Furthermore, $\text{dom } A(p, q)$ is dense in $V_0(p)$ and $A(p, q)$ generates an analytic semigroup $T(t; p, q)$ on $H(p)$. Mild solutions of the abstract evolution equations (2.5), (2.6) are given in terms of this semigroup by

$$(3.2) \quad u(t; p, q) = T(t; p, q)u_0(p, q) + \int_0^t T(t-s; p, q)F(s; p, q) ds,$$

where $u_0(p, q)$ and $F(\cdot; p, q)$ are given in $H(p)$ and are understood in the sense described in the last section. In addition, for $\lambda \geq \lambda_0$, the resolvent operator $R_\lambda(A(p, q))$ exists and is a bounded linear operator on $H(p)$.

Remark 3.3. It will often be more convenient to refer to a slightly different statement of the continuity (in p) of the $H(p)$ inner product and the sesquilinear form $\sigma(p, q)$ than is given above. Specifically, from the assumptions about γ and γ_e , we observe that for any $p_1, p_2 \in \mathcal{P}$ and arbitrary $z_1 \in H(p_1)$, $z_2 \in H(p_2)$,

$$\begin{aligned} & \left| \langle \gamma(\tilde{p}, p_1)z_1, \gamma(\tilde{p}, p_2)z_2 \rangle_{\tilde{p}} - \langle \gamma(p, p_1)z_1, \gamma(p, p_2)z_2 \rangle_p \right| \\ &= \left| \langle \gamma_t(\tilde{p})(\gamma_e(p_1)z_1), \gamma_t(\tilde{p})(\gamma_e(p_2)z_2) \rangle_{\tilde{p}} - \langle \gamma_t(p)(\gamma_e(p_1)z_1), \gamma_t(p)(\gamma_e(p_2)z_2) \rangle_p \right| \\ &\leq K^2 C_p(\tilde{p}) \|z_1\|_{p_1} \|z_2\|_{p_2}. \end{aligned}$$

Similarly, for any $v_1 \in V_0(p_1)$ and $v_2 \in V_0(p_2)$,

$$\begin{aligned} & \left| \sigma(\tilde{p}, q)(\gamma(\tilde{p}, p_1)v_1, \gamma(\tilde{p}, p_2)v_2) - \sigma(p, q)(\gamma(p, p_1)v_1, \gamma(p, p_2)v_2) \right| \\ &\leq K^2 C_{p,q}(\tilde{p}) \|v_1\|_{p_1} \|v_2\|_{p_2}. \end{aligned}$$

We now turn to the construction of approximation spaces and define the (non-parametrized) finite-dimensional spaces $H^N \subseteq V_0$, for each $N = 1, 2, \dots$. For each $p \in \mathcal{P}$, the parametrized approximation spaces $H^N(p)$ are constructed from H^N using the (theoretical) truncation map $\gamma_t(p)$, $H^N(p) \equiv \gamma_t(p)(H^N)$; thus the $H^N(p)$ are finite-dimensional subspaces of $H(p)$ satisfying $H^N(p) \subseteq V_0(p)$. We note that, in the general case, we do not have $H^N(p)$ isomorphic to H^N (because we need not have $\gamma_t(p)$ an isomorphism on H^N ; i.e., $\ker \gamma_t(p) \cap H^N \neq \{0\}$ in general); it is, however, usually the case that $\gamma_t(p)$ is an isomorphism off of its kernel (see, for example, [19]). With this in mind, we make the following standing assumptions about $H^N(p)$.

Hypothesis 3 (Properties of $H^N(p)$).

(a) *Invertibility of $\gamma_t(p)$ on $H^N(p)$:* For each N , the restriction of $\gamma_t(p)$ to $H^N \cap (\ker \gamma_t(p))^\perp$ is an isomorphism with uniformly (in p) bounded inverse; i.e., there exists a constant $c(N) > 0$ such that

$$\|\gamma_t(p)v\|_p \geq c(N)\|v\|$$

for all $p \in \mathcal{P}$ and all $v \in H^N \cap (\ker \gamma_t(p))^\perp$. (Here “ \perp ” denotes the orthogonal complement in the V topology.)

(b) *Uniform (in p) approximation properties:* For each $p \in \mathcal{P}$ and $v_0 \in V_0(p)$, there exists $\hat{v}_0^N(p) \in H^N(p)$ such that

$$\|v_0 - \hat{v}_0^N(p)\|_p \rightarrow 0 \quad \text{as } N \rightarrow \infty,$$

where the rate of convergence does not depend explicitly on p , but may increase monotonically with $\|v_0\|_p$.

Remark 3.4. It should be noted that if the nonparametrized spaces H^N satisfy an approximation estimate like that given in assumption (b) of Hypothesis 3, then, provided that $\gamma_t(p)$ maps V_0 onto $V_0(p)$, the *uniform* (in p) approximation properties required for $H^N(p)$ follow immediately.

Remark 3.5. From the assumptions made in this section, we obtain essential results about $P^N(p)$ and $P_V^N(p)$, the two orthogonal (with regard to the $H(p)$ and $V(p)$ inner products, respectively) projection operators, $P^N(p) : H(p) \rightarrow H^N(p)$, $P_V^N(p) : V(p) \rightarrow H^N(p)$. We first use the uniform (in p) approximation properties of $H^N(p)$ to argue that, given $\epsilon > 0$, $p \in \mathcal{P}$, and $v_0 \in V_0(p)$, there exists $\mathcal{N} > 0$ (where \mathcal{N} depends on ϵ and $K^2\|v_0\|_p$, an upper bound for $\|\gamma(\tilde{p}, p)v_0\|_{\tilde{p}}$), such that for any integer $N \geq \mathcal{N}$ and *arbitrary* $\tilde{p} \in \mathcal{P}$, we have

$$\begin{aligned} \|(P_V^N(\tilde{p}) - I)\gamma(\tilde{p}, p)v_0\|_{\tilde{p}} &\leq \|\hat{v}_0^N(\tilde{p}) - \gamma(\tilde{p}, p)v_0\|_{\tilde{p}} \\ &< \epsilon. \end{aligned}$$

Because $V_0(p)$ is dense in $H(p)$, we get a similar estimate for P^N : given arbitrary $z \in H(p)$ and $\tilde{p} \in \mathcal{P}$, we have $|(P^N(\tilde{p}) - I)\gamma(\tilde{p}, p)z|_{\tilde{p}} < \epsilon$, for N sufficiently large (the size of N does not depend on \tilde{p}).

In addition, appealing to (c) in Hypothesis 1, in which it is assumed that $\pi(\tilde{p}, p)$ approximates $\gamma(\tilde{p}, p)$ for \tilde{p} close to p , we are able to make a similar statement about these projections when the “data” space-changing operator π is used instead of γ . That is, for arbitrary $p^N \rightarrow p$ in \mathcal{P} and $v_0 \in V_0(p)$, an application of the triangle inequality yields

$$\begin{aligned} &\|(P_V^N(p^N) - I)\pi(p^N, p)v_0\|_{p^N} \\ &\leq \|P_V^N(p^N)(\pi(p^N, p) - \gamma(p^N, p))v_0\|_{p^N} + \|(P_V^N(p^N) - I)\gamma(p^N, p)v_0\|_{p^N} \\ &\quad + \|(\gamma(p^N, p) - \pi(p^N, p))v_0\|_{p^N} \end{aligned}$$

so that for any $v_0 \in V_0(p)$,

$$(3.3) \quad \|(P_V^N(p^N) - I)\pi(p^N, p)v_0\|_{p^N} \rightarrow 0, \quad \text{as } N \rightarrow \infty,$$

and similarly, for arbitrary $z \in H(p)$,

$$(3.4) \quad |(P^N(p^N) - I)\pi(p^N, p)z|_{p^N} \rightarrow 0, \quad \text{as } N \rightarrow \infty.$$

In fact, we also obtain similar estimates for $P_V^M(p^N)$ and $P^M(p^N)$ (the projections from $V(p^N)$ and $H(p^N)$, respectively, into $H^M(p^N)$) where M is *fixed*. We first take $p, \tilde{p} \in \mathcal{P}$ and observe that for arbitrary $v^M(p)$ in $H^M(p)$,

$$\|(P_V^M(\tilde{p}) - I)\gamma(\tilde{p}, p)v^M(p)\|_{\tilde{p}} \leq \|\gamma_t(\tilde{p})v_p^M - \gamma_t(\tilde{p})\gamma_e(p)v^M(p)\|_{\tilde{p}},$$

where, using assumption (a) of Hypothesis 3, v_p^M is uniquely given in $H^M \cap (\ker \gamma_t(p))^\perp$, satisfying $v^M(p) = \gamma_t(p)v_p^M$. It thus follows that

$$\begin{aligned} \|(P_V^M(\tilde{p}) - I)\gamma(\tilde{p}, p)v^M(p)\|_{\tilde{p}} &\leq \|(\gamma_t(\tilde{p}) - \gamma_t(\tilde{p})\gamma_e(p)\gamma_t(p))v_p^M\|_{\tilde{p}} \\ &\leq C_\gamma(p, \tilde{p})\|v_p^M\| \\ &\leq C_\gamma(p, \tilde{p})\|v^M(p)\|_p/c(M), \end{aligned}$$

where part (iii) of Hypothesis 1(b), and assumption (a) of Hypothesis 3 have been used. We may argue similarly for $P^M(\tilde{p})$; we thus find there exists a constant

$K_1(M) > 0$ such that, for arbitrary $p, \tilde{p} \in \mathcal{P}$,

$$(3.5) \quad \|(P_V^M(\tilde{p}) - I)\gamma(\tilde{p}, p)v^M(p)\|_{\tilde{p}} \leq K_1(M)C_\gamma(p, \tilde{p})\|v^M(p)\|_p,$$

$$(3.6) \quad |(P^M(\tilde{p}) - I)\gamma(\tilde{p}, p)v^M(p)|_{\tilde{p}} \leq K_1(M)C_\gamma(p, \tilde{p})\|v^M(p)\|_p$$

and therefore, for each fixed M and $(p^N, q^N) \rightarrow (p, q)$,

$$(3.7) \quad \|(P_V^M(p^N) - I)\gamma(p^N, p)v^M(p)\|_{p^N} \rightarrow 0 \quad \text{as } N \rightarrow \infty,$$

$$(3.8) \quad |(P^M(p^N) - I)\gamma(p^N, p)v^M(p)|_{p^N} \rightarrow 0 \quad \text{as } N \rightarrow \infty,$$

for arbitrary $v^M(p) \in H^M(p)$. Because $\pi(p^N, p)$ approximates $\gamma(p^N, p)$ for N sufficiently large, we may also write statements like the above using instead the operator $\pi(p^N, p)$ in each case.

As was indicated in the last section, we define for each $(p, q) \in \mathcal{P} \times \mathcal{Q}$, approximations of Galerkin type through the definition of operators $A^N(p, q) : H^N(p) \rightarrow H^N(p)$ (defined by the restriction of $\sigma(p, q)$ to $H^N(p) \times H^N(p)$), and define corresponding semigroups $T^N(t; p, q)$. Associated then with the ordinary differential equation system (2.8), (2.9) in $u^N(t; p, q)$ on $H^N(p)$ is a set of approximating equations for (3.2),

$$(3.9) \quad u^N(t; p, q) = T^N(t; p, q)P^N(p)u_0(p, q) + \int_0^t T^N(t-s; p, q)P^N(p)F(s; p, q) ds,$$

or, more generally, for given M and $(p^N, q^N) \in \mathcal{P} \times \mathcal{Q}$, we have the equation on $H^M(p^N)$,

$$(3.10) \quad \begin{aligned} u^M(t; p^N, q^N) &= T^M(t; p^N, q^N)P^M(p^N)u_0(p^N, q^N) \\ &\quad + \int_0^t T^M(t-s; p^N, q^N)P^M(p^N)F(s; p^N, q^N) ds. \end{aligned}$$

Our ultimate goal is to obtain the continuous dependence and convergence results stated in the last section, and these obviously cannot be accomplished unless the parameter-dependent initial condition $u_0(p, q)$ and perturbation term $F(\cdot; p, q)$ are known to depend continuously on these parameters. We make these assumptions here, although the hypotheses will not be used until the final convergence theorems are stated (Theorems 3.3 and 3.6).

Hypothesis 4 (Continuity of data). Given $(p^N, q^N) \in \mathcal{P} \times \mathcal{Q}$ satisfying $(p^N, q^N) \rightarrow (p, q) \in \mathcal{P} \times \mathcal{Q}$ as $N \rightarrow \infty$, the initial data u_0 and external force F satisfy

$$\begin{aligned} |u_0(p^N, q^N) - \pi(p^N, p)u_0(p, q)|_{p^N} &\rightarrow 0, \\ |F(s; p^N, q^N) - \pi(p^N, p)F(s; p, q)|_{p^N} &\rightarrow 0, \end{aligned}$$

for each $s \in (0, T)$, as $N \rightarrow \infty$.

Appealing to this assumption, there is no loss in generality in using the following equation for $u^M(\cdot; p^N, q^N)$, in place of (3.10):

$$(3.11) \quad \begin{aligned} u^M(t; p^N, q^N) &= T^M(t; p^N, q^N)P^M(p^N)\pi(p^N, p)u_0(p, q) \\ &\quad + \int_0^t T^M(t-s; p^N, q^N)P^M(p^N)\pi(p^N, p)F(s; p, q) ds. \end{aligned}$$

In keeping with the development in [3], the required continuous dependence and convergence arguments will be established through the use of a version of the Trotter–Kato theorem [23] in “resolvent convergence form” (i.e., one in which the convergence of $R_\lambda(A^M(p^N, q^N))$ is required). It is the use of this form of the theorem, rather than the usual “operator convergence form” (where convergence of $A^M(p^N, q^N)$ must be demonstrated) that allows significant improvement in the types of assumptions required to obtain a parameter convergence theory (specifically, assumptions of compactness on $\mathcal{P} \times \mathcal{Q}$ may be made with regard to a less restrictive topology, this achieved without requiring increased regularity of solutions $u(t; p, q)$ [3]). The exact form of the Trotter–Kato theorem used here [5] differs from that employed in [3] due to the need for multiple state spaces.

LEMMA 3.1 (Trotter–Kato theorem). *For each $N = 1, 2, \dots$, let $(X, |\cdot|)$ and $(X^N, |\cdot|_N)$ be Banach spaces and $\mathcal{B}^N : X \rightarrow X^N$ a bounded linear operator. Let A^N , A be infinitesimal generators of C_0 semigroups $S^N(t), S(t)$ on X^N , X respectively. Assume the following:*

- (i) *There exists $K \geq 0$ such that $|\mathcal{B}^N| \leq K$ for every N ;*
- (ii) *There exist constants ω and M such that $|S^N(t)| \leq Me^{\omega t}$ for each N ;*
- (iii) *There exists $\lambda \in \rho(A) \cap \bigcap_{N=1}^\infty \rho(A^N)$ such that $\operatorname{Re} \lambda > \omega$ and for each $x \in X$,*

$$|R_\lambda(A^N)\mathcal{B}^N x - \mathcal{B}^N R_\lambda(A)x|_N \rightarrow 0, \quad \text{as } N \rightarrow \infty.$$

(where $\rho(A)$, $\rho(A^N)$ are the resolvent sets for A , A^N , respectively). Then, for every $x \in X$,

$$|S^N(t)\mathcal{B}^N x - \mathcal{B}^N S(t)x|_N \rightarrow 0, \quad \text{as } N \rightarrow \infty,$$

uniformly in $t \in [0, T]$.

Before making direct application of this result, we first obtain some preliminary estimates for the resolvent operators, which will be used to construct the convergence arguments, $\|u^N(t; p^N, q^N) - \pi(p^N, p)u(t; p, q)\|_{p^N} \rightarrow 0$ as $N \rightarrow \infty$ (i.e., step 2 of the last section).

THEOREM 3.2. *Let $(p^N, q^N) \rightarrow (p, q)$ in $\mathcal{P} \times \mathcal{Q}$. Then for $\lambda = \lambda_0$ (λ_0 given in assumption (c) of Hypothesis 2) and arbitrary $z \in H(p)$, we have*

$$\|R_\lambda(A^N)P^N\pi^N z - \pi^N R_\lambda(A)z\|_N \rightarrow 0,$$

as $N \rightarrow \infty$. Here we have introduced the abbreviated notation $A^N \equiv A^N(p^N, q^N)$, $A \equiv A(p, q)$, $P^N \equiv P^N(p^N)$, $\pi^N \equiv \pi(p^N, p)$, and $\|\cdot\|_N \equiv \|\cdot\|_{p^N}$.

Proof. Because the choice $\lambda = \lambda_0$ implies that $\lambda \in \rho(A) \cap \bigcap_{N=1}^\infty \rho(A^N)$, there is no difficulty in defining, for arbitrary $z \in H(p)$, the quantities $w = w(p, q) \in V_0(p)$ and $w^N = w^N(p^N, q^N) \in H^N(p^N)$ by

$$w = R_\lambda(A)z \quad \text{and} \quad w^N = R_\lambda(A^N)P^N\pi^N z,$$

where our goal is to prove that $\|w^N - \pi^N w\|_N \rightarrow 0$ as $N \rightarrow \infty$. Using the estimate

$$\|w^N - \pi^N w\|_N \leq \|w^N - P_V^N \pi^N w\|_N + \|(P_V^N - I)\pi^N w\|_N,$$

it is clear, in view of estimates for P_V^N given in (3.3), that we need only prove convergence of the first term, or that $\|w^N - \hat{w}^N\|_N \rightarrow 0$, where $\hat{w}^N \equiv P_V^N \pi^N w \in H^N(p^N)$.

In the subsequent discussion, we simplify notation by defining $z^N \equiv w^N - \hat{w}^N \in H^N(p^N)$, $\gamma^N \equiv \gamma(p^N, p)$, $\langle \cdot, \cdot \rangle_N \equiv \langle \cdot, \cdot \rangle_{p^N}$, and $|\cdot|_N \equiv |\cdot|_{p^N}$. It follows from the uniform coercivity of $\sigma(p^N, q^N)$ that

$$\begin{aligned} c_1 \|z^N\|_N^2 &\leq \sigma(p^N, q^N)(z^N, z^N) + \lambda |z^N|_N^2 \\ &= \sigma(p^N, q^N)(w^N, z^N) - \sigma(p^N, q^N)(\hat{w}^N, z^N) + \lambda |z^N|_N^2 \\ &\leq T_1^N + T_2^N + T_3^N, \end{aligned}$$

where

$$\begin{aligned} T_1^N &\equiv \sigma(p^N, q^N)(w^N, z^N) - \sigma(p, q)(w, \gamma(p, p^N)z^N) + \lambda |z^N|_N^2, \\ T_2^N &\equiv |\sigma(p, q)(w, \gamma(p, p^N)z^N) - \sigma(p^N, q)(\gamma^N w, z^N)|, \\ T_3^N &\equiv |\sigma(p^N, q)(\gamma^N w, z^N) - \sigma(p^N, q^N)(\hat{w}^N, z^N)|. \end{aligned}$$

Considering first T_1^N and making use of the fact that w^N and z^N belong to $H^N(p^N)$, while $w \in \text{dom} A$ and $\gamma(p, p^N)z^N \in V_0(p)$, we may appeal to properties of the operators A^N and A to obtain

$$\begin{aligned} T_1^N &= \langle (\lambda I - A^N - \lambda I)w^N, z^N \rangle_N - \langle (\lambda I - A - \lambda I)w, \gamma(p, p^N)z^N \rangle_p + \lambda |z^N|_N^2 \\ (3.12) \quad &= \{ \langle P^N \pi^N z, z^N \rangle_N - \langle z, \gamma(p, p^N)z^N \rangle_p \} \\ &\quad - \lambda \{ \langle w^N, z^N \rangle_N - \langle w, \gamma(p, p^N)z^N \rangle_p - |z^N|_N^2 \}. \end{aligned}$$

Adding and subtracting terms, and making liberal use of the identities

$$z^N = \gamma(p^N, p^N)z^N, \quad z = \gamma(p, p)z, \quad \gamma^N \equiv \gamma(p^N, p),$$

we observe that

$$\begin{aligned} T_1^N &\leq |\langle \pi^N z, z^N \rangle_N - \langle \gamma^N z, z^N \rangle_N| \\ &\quad + |\langle \gamma(p^N, p)z, \gamma(p^N, p^N)z^N \rangle_N - \langle \gamma(p, p)z, \gamma(p, p^N)z^N \rangle_p| \\ &\quad - \lambda \{ \langle w^N - \hat{w}^N, z^N \rangle_N + |\langle \hat{w}^N, z^N \rangle_N - \langle \gamma^N w, z^N \rangle_N| \\ &\quad + |\langle \gamma(p^N, p)w, \gamma(p^N, p^N)z^N \rangle_N - \langle \gamma(p, p)w, \gamma(p, p^N)z^N \rangle_p| - |z^N|_N^2 \}, \end{aligned}$$

where we have used the orthogonality of $P^N : H(p^N) \rightarrow H^N(p^N)$ in the first term. Appealing to the Cauchy-Schwarz inequality and the continuity property of the $H(p)$ inner product, we conclude that

$$\begin{aligned} T_1^N &\leq \{ |(\pi^N - \gamma^N)z|_N |z^N|_N \\ &\quad + K^2 C_p(p^N) |z|_p |z^N|_N \} + |\lambda| \{ |\hat{w}^N - \gamma^N w|_N |z^N|_N + K^2 C_p(p^N) |w|_p |z^N|_N \} \\ &= \{ |(\pi^N - \gamma^N)z|_N + K^2 C_p(p^N) (|z|_p + |\lambda| |w|_p) + k |\lambda| \|\hat{w}^N - \gamma^N w\|_N \} k \|z^N\|_N. \end{aligned}$$

In addition, from the continuity (in p) of $\sigma(p, q)$, we find that

$$\begin{aligned} T_2^N &= |\sigma(p, q)(\gamma(p, p)w, \gamma(p, p^N)z^N) - \sigma(p^N, q)(\gamma(p^N, p)w, \gamma(p^N, p^N)z^N)| \\ &\leq K^2 C_{p,q}(p^N) \|w\|_p \|z^N\|_N, \end{aligned}$$

while the (uniform in p) q -continuity of the sesquilinear form, coupled with uniform V_0 -boundedness yields the remaining estimate

$$\begin{aligned} T_3^N &\leq |\sigma(p^N, q)(\gamma^N w, z^N) - \sigma(p^N, q^N)(\gamma^N w, z^N)| + |\sigma(p^N, q^N)(\gamma^N w - \hat{w}^N, z^N)| \\ &\leq (C_q(q^N) K^2 \|w\|_p + c_2 \|\gamma^N w - \hat{w}^N\|_N) \|z^N\|_N. \end{aligned}$$

Combining the three inequalities, we conclude that

$$c_1 \|z^N\|_N \leq K_0 \{ C_p(p^N) + C_{p,q}(p^N) + C_q(q^N) + |(\pi^N - \gamma^N)z|_N + \|\gamma^N w - \hat{w}^N\|_N \},$$

for $K_0 > 0$ a constant independent of N . Finally, from the observation that

$$\|\gamma^N w - \hat{w}^N\|_N \leq \|(\gamma^N - \pi^N)w\|_N + \|\pi^N w - P_V^N \pi^N w\|_N,$$

it follows that $\|\gamma^N w - \hat{w}^N\|_N \rightarrow 0$, and thus, $\|z^N\|_N \rightarrow 0$ as $N \rightarrow \infty$. \square

Using these resolvent estimates, we appeal to the Trotter–Kato theorem to argue convergence (in the proper sense) of $u^N(t; p^N, q^N)$ to $u(t; p, q)$. We note that at this point an easy application of this theorem yields the convergence of semigroups in the $H(p^N)$ norm

$$(3.13) \quad |T^N(t)P^N\pi^N z - P^N\pi^N T(t)z|_N \rightarrow 0,$$

as $N \rightarrow \infty$, uniformly in $t \in [0, T]$ (where we simplify notation here and throughout by defining $T^N(t) \equiv T^N(t; p^N, q^N)$ and $T(t) \equiv T(t; p, q)$) This statement of semigroup convergence follows if we apply the Trotter–Kato theorem using $X = (H(p), |\cdot|_p)$, and $X^N = (H^N(p^N), |\cdot|_N)$, $\mathcal{B}^N = P^N\pi^N$, $S^N(t) = T^N(t; p^N, q^N)$, and $S(t) = T(t; p, q)$. Using these definitions, conditions (i) and (ii) of the theorem are easily verified, while for (iii), we make the calculation

$$\begin{aligned} & |R_\lambda(A^N)P^N\pi^N z - P^N\pi^N R_\lambda(A)z|_N \\ & \leq k \|R_\lambda(A^N)P^N\pi^N z - \pi^N R_\lambda(A)z\|_N + |(P^N - I)\pi^N R_\lambda(A)z|_N \end{aligned}$$

and use Theorem 3.2 and the statement of projection convergence in (3.4) to confirm that we have satisfied the required convergence of resolvent operators. We thus obtain the semigroup convergence in (3.13), and, in fact, may use this result with $|(P^N - I)\pi^N T(t)z|_N \rightarrow 0$ to establish that

$$|T^N(t)P^N\pi^N z - \pi^N T(t)z|_N \rightarrow 0.$$

From this estimate, it easily follows that

$$|u^N(t; p^N, q^N) - \pi^N u(t; p, q)|_N \rightarrow 0,$$

as $N \rightarrow \infty$, uniformly in $t \in [0, T]$.

However, as in [3], it is our goal to establish these conclusions in the stronger norm $\|\cdot\|_N$. Convergence in the V topology is realized in the next theorem.

THEOREM 3.3. *Under the same conditions as Theorem 3.2, we find that for arbitrary $z \in H(p)$,*

$$(3.14) \quad \|T^N(t)P^N\pi^N z - \pi^N T(t)z\|_N \rightarrow 0,$$

and thus,

$$(3.15) \quad \|u^N(t; p^N, q^N) - \pi^N u(t; p, q)\|_N \rightarrow 0,$$

as $N \rightarrow \infty$, where the convergence in each case is at a rate uniform in $t > 0$ in compact subintervals.

Proof. The basic arguments that we pursue differ little from those found in the proof of Theorem 2.3 in [3]. We briefly summarize the results in that theorem as they apply to the case of parameter-dependent state spaces.

We will make use of bounds on resolvents and semigroups nearly identical to those employed by Banks and Ito, except for obvious changes in norms and spaces. These estimates are derived in [3] (specifically, we refer to inequalities (2.8), (2.11), and (2.12) of that reference) using results found in Tanabe [26, Lemma 6.1, Chap. 3]. First, for $\lambda \geq \lambda_0$ (where λ_0 is taken here and in [3] to be zero, without loss of generality) and arbitrary $f_0(p^N) \in V_0(p^N)$ we have the resolvent bound

$$(3.16) \quad \|R_\lambda(A^N)f_0(p^N)\|_N \leq c|f_0(p^N)|_N/|\lambda|^{1/2},$$

where $c \geq 0$ does not depend on N . The parameter-dependent analogues of (2.11) and (2.12) in [3] are the estimates for semigroups $T^N(t)$, valid for all $v^N \in H^N(p^N)$ and some \mathcal{M} independent of N

$$(3.17) \quad \|T^N(t)v^N\|_N \leq \mathcal{M}e^{\lambda_0 t}\|v^N\|_N.$$

and, for $t > 0$ and $\tilde{\mathcal{M}}$ independent of N ,

$$(3.18) \quad \|T^N(t)v^N\|_N \leq \tilde{\mathcal{M}}e^{\lambda_0 t}t^{-1/2}|v^N|_N.$$

Finally, for every $z \in H(p)$, we have $T(t)z \in V_0(p)$ for $t > 0$, and, using density of $V_0(p)$ in $H(p)$ (along with estimates (3.16) above and (2.10) of [3]), the bound

$$(3.19) \quad \|T(t)z\|_p \leq \tilde{\mathcal{M}}e^{\lambda_0 t}t^{-1/2}|z|_p$$

is determined (this estimate is also used in [3], in the last paragraph of the proof of Theorem 2.3 in that reference).

We turn now to the proof of (3.14), first establishing the intermediate result,

$$(3.20) \quad \|T^N(t)P_V^N\pi^N v_0 - P_V^N\pi^N T(t)v_0\|_N \rightarrow 0,$$

for arbitrary $v_0 \in V_0(p)$. Again we use the Trotter–Kato theorem, this time with $X = (V_0(p), \|\cdot\|_p)$, $X^N = (H^N(p^N), \|\cdot\|_N)$, and $\mathcal{B}^N = P_V^N\pi^N$. Because $\|P_V^N\pi^N\| \leq K$ for all N , (i) in that theorem is valid, and we may appeal to (3.17) to obtain condition (ii). Furthermore, we argue the resolvent convergence in (iii) by taking steps similar to those in [3]: here we use the triangle inequality and the bound in (3.16) (with $f_0(p^N) = (P_V^N - P^N)\pi^N v_0$) and find that

$$(3.21) \quad \begin{aligned} & \|R_\lambda(A^N)P_V^N\pi^N v_0 - P_V^N\pi^N R_\lambda(A)v_0\|_N \\ & \leq c|(P_V^N - P^N)\pi^N v_0|_N/|\lambda|^{1/2} + \|R_\lambda(A^N)P^N\pi^N v_0 - \pi^N R_\lambda(A)v_0\|_N \\ & \quad + \|(I - P_V^N)\pi^N R_\lambda(A)v_0\|_N. \end{aligned}$$

If (3.3) and (3.4) are used (recalling that $R_\lambda(A)v_0 \in V_0(p)$), it may be seen that the first and third terms in the above expression converge to zero, while convergence of the remaining term is due to Theorem 3.2. The Trotter–Kato theorem may therefore be applied to conclude the convergence in (3.20) is valid at a rate uniform in $t \geq 0$ in compact subintervals.

To complete the proof, we wish to establish that $\|T^N(t)P^N\pi^N z - \pi^N T(t)z\|_N \rightarrow 0$ for arbitrary $z \in H(p)$. Density of $V_0(p)$ in $H(p)$ guarantees the existence of $v_0(p) \in V_0(p)$ arbitrarily close to $z \in H(p)$; an application of the triangle inequality then yields

$$(3.22) \quad \begin{aligned} & \|T^N(t)P^N\pi^N z - \pi^N T(t)z\|_N \\ & \leq \|T^N(t)(P^N\pi^N z - P_V^N\pi^N v_0)\|_N + \|T^N(t)P_V^N\pi^N v_0 - P_V^N\pi^N T(t)v_0\|_N \\ & \quad + \|P_V^N\pi^N T(t)(v_0 - z)\|_N + \|(P_V^N - I)\pi^N T(t)z\|_N. \end{aligned}$$

Using the estimate in (3.18), the first term on the right side may be bounded, for $t > 0$, by

$$\begin{aligned} & \|T^N(t)(P^N\pi^N z - P_V^N\pi^N v_0)\|_N \\ & \leq \tilde{\mathcal{M}}e^{\lambda_0 t}t^{-1/2}(|P^N\pi^N(z - v_0)|_N + |(P^N - P_V^N)\pi^N v_0|_N), \end{aligned}$$

while the third term in (3.22) satisfies

$$\|P_V^N\pi^N T(t)(v_0 - z)\|_N \leq K^2 \tilde{\mathcal{M}}e^{\lambda_0 t}t^{-1/2}|v_0 - z|_p.$$

Finally, using the convergence properties of P_V^N in the last term in (3.22) (using $T(t)z \in V_0(p)$ for $t > 0$), we see that all terms in (3.22) may be made arbitrarily small for $t > 0$ and N sufficiently large. The desired $V(p^N)$ convergence of semigroups given in the statement of the theorem (uniform in $t > 0$ in compact subintervals) then follows immediately. Using this convergence, we easily obtain

$$\|u^N(t; p^N, q^N) - \pi^N u(t; p, q)\|_N \rightarrow 0 \quad \text{as } N \rightarrow \infty,$$

at a rate that is also uniform in compact subintervals of $t > 0$. \square

We now turn to a consideration of the continuous dependence of $u^M(\cdot; p, q)$ on (p, q) ; to this end, we construct arguments that are actually only slight variations of those used in the first two theorems (once we prove some initial estimates regarding the dependence of the projection operators on the parameter p). The needed preliminary results are summarized in the following lemma.

LEMMA 3.4. *Let $M = 1, 2, \dots$, be fixed. Given any $(p^N, q^N) \in \mathcal{P} \times \mathcal{Q}$ with $(p^N, q^N) \rightarrow (p, q) \in \mathcal{P} \times \mathcal{Q}$ and arbitrary $z \in H(p)$, we have*

$$(3.23) \quad \|P^M(p^N)\pi^N z - \pi^N P^M(p)z\|_N \rightarrow 0$$

as $N \rightarrow \infty$.

Proof. We establish the desired convergence via an examination of the continuity properties (with respect to the parameter p) of basis elements for $H^M(p)$. Let $\{B_i^M\}_{i=1}^n$ denote a basis for the fixed space H^M , $n = n(M, p)$, which satisfies $\text{span}\{B_i^M\}_{i=1}^l = H^M \cap (\ker \gamma_t(p))^\perp$ and $\text{span}\{B_i^M\}_{i=l+1}^n = (\ker \gamma_t(p)) \cap H^M$, for $1 \leq l \leq n$. (We note that (b) of Hypothesis 3 guarantees that $H^M \cap (\ker \gamma_t(p))^\perp$ is nonempty for every p ; we assume for the purposes of this proof, without loss of generality, that $(\ker \gamma_t(p)) \cap H^M$ is nonempty as well.) Defining $B_i^M(p) \equiv \gamma_t(p)B_i^M$, $B_i^M(p^N) \equiv \gamma_t(p^N)B_i^M$ for $i = 1, \dots, n$, we note that the corresponding elements “get close” in the sense that

$$\begin{aligned} \|B_i^M(p^N) - \gamma^N B_i^M(p)\|_N & \equiv \|\gamma_t(p^N)B_i^M - \gamma_t(p^N)\gamma_e(p)\gamma_t(p)B_i^M\|_N \\ & \leq C_\gamma(p, p^N)\|B_i^M\|, \quad \text{for } i = 1, \dots, n, \end{aligned}$$

which may be seen to converge to zero as $N \rightarrow \infty$. The conditions on B_i^M imply that $\{B_i^M(p)\}_{i=1}^l$ is a basis for $H^M(p)$, and, because these elements are linearly independent, we have that their Gramian [22, p. 110], $g(B_1^M(p), \dots, B_l^M(p)) \equiv \det(\langle B_i^M(p), B_j^M(p) \rangle_p)_{i,j=1}^l$ is strictly positive. It is also clear, from the definition of $B_i^M(p^N)$ and the p -continuity of the $H(p)$ inner product, that for N sufficiently large, we have $g(B_1^M(p^N), \dots, B_l^M(p^N))$ strictly positive as well; thus these elements are also linearly independent, although they need not span $H^M(p^N)$. There is no loss of generality, however in assuming that these elements, $\{B_i^M(p^N)\}_{i=1}^l$, augmented with the set $\{B_i^M(p^N)\}_{i=l+1}^m$, $l \leq m$, form a basis for $H^M(p^N)$.

We next construct orthonormal sets $\{G_i^M(p)\}_{i=1}^l$ for $H^M(p)$ and $\{G_i^M(p^N)\}_{i=1}^l$ for $H^M(p^N)$ (orthonormal in $\langle \cdot, \cdot \rangle_p$ and $\langle \cdot, \cdot \rangle_N$, respectively) using the Gram–Schmidt process on the first l basis elements (taken in order) for each space. Because the inner product $\langle B_i^M(p^N), B_j^M(p^N) \rangle_N$ approximates $\langle B_i^M(p), B_j^M(p) \rangle_p$ for $1 \leq i, j \leq l$ and N sufficiently large, it is not difficult to show (using induction) that the Gram–Schmidt process generates orthonormal elements that preserve many of the properties of the original elements; for example,

$$\|G_i^M(p^N) - \gamma^N G_i^M(p)\|_N \rightarrow 0 \quad \text{as } N \rightarrow \infty.$$

For the remaining $H^M(p^N)$ basis elements $\{B_i^M(p^N)\}_{i=l+1}^m$, we again use the Gram–Schmidt process, this time constructing elements that are not normalized; that is, for $i = l+1, \dots, m$, we define

$$G_i^M(p^N) \equiv B_i^M(p^N) - \sum_{j=1}^{i-1} \langle B_i^M(p^N), G_j^M(p^N) \rangle_N G_j^M(p^N).$$

We observe that, for $i = l+1, \dots, m$, each $G_i^M(p^N)$ may be approximated by

$$\gamma^N B_i^M(p) - \sum_{j=1}^{i-1} \langle B_i^M(p), G_j^M(p) \rangle_p \gamma^N G_j^M(p),$$

but that this expression equals zero because that $B_i^M(p) \equiv \gamma_t(p) B_i^M = 0$ for each $i \geq l+1$. Thus, we have

$$\|G_i^M(p^N)\|_N \rightarrow 0 \quad \text{as } N \rightarrow \infty,$$

for $i = l+1, \dots, m$.

To argue the convergence in (3.23), we make preliminary calculations using the operator γ^N instead of π^N (for which the statement of (3.23) is given). For arbitrary $z \in H(p)$,

$$\begin{aligned} P^M(p)z &= \sum_{i=1}^l \langle z, G_i^M(p) \rangle_p G_i^M(p), \\ P^M(p^N)\gamma^N z &= \sum_{i=1}^l \langle \gamma^N z, G_i^M(p^N) \rangle_N G_i^M(p^N) + \sum_{i=l+1}^m \frac{\langle \gamma^N z, G_i^M(p^N) \rangle_N}{|G_i^M(p^N)|_N} G_i^M(p^N) \end{aligned}$$

so that

$$\begin{aligned} &\|P^M(p^N)\gamma^N z - \gamma^N P^M(p)z\|_N \leq \\ &\leq \sum_{i=1}^l \|\langle \gamma^N z, G_i^M(p^N) \rangle_N G_i^M(p^N) - \langle z, G_i^M(p) \rangle_p \gamma^N G_i^M(p)\|_N \\ &\quad + \sum_{i=l+1}^m \left\| \frac{\langle \gamma^N z, G_i^M(p^N) \rangle_N}{|G_i^M(p^N)|_N} G_i^M(p^N) \right\|_N \\ &\leq \sum_{i=1}^l |\langle \gamma^N z, G_i^M(p^N) \rangle_N| \|G_i^M(p^N) - \gamma^N G_i^M(p)\|_N \end{aligned}$$

$$\begin{aligned}
& + \sum_{i=1}^l |\langle \gamma^N z, G_i^M(p^N) - \gamma^N G_i^M(p) \rangle_N| \|\gamma^N G_i^M(p)\|_N \\
& + \sum_{i=1}^l |\langle \gamma^N z, \gamma^N G_i^M(p) \rangle_N - \langle z, G_i^M(p) \rangle_p| \|\gamma^N G_i^M(p)\|_N \\
& + \sum_{i=l+1}^m |\gamma^N z|_N \|G_i^M(p^N)\|_N \\
& \leq K^2 |z|_p \sum_{i=1}^l k \|G_i^M(p^N)\|_N \|G_i^M(p^N) - \gamma^N G_i^M(p)\|_N \\
& + |z|_p \sum_{i=1}^l k (K^2 \|G_i^M(p^N) - \gamma^N G_i^M(p)\|_N + C_p(p^N) \|G_i^M(p)\|_p) K^2 \|G_i^M(p)\|_p \\
& + K^2 |z|_p \sum_{i=l+1}^m \|G_i^M(p^N)\|_N,
\end{aligned}$$

which converges to 0 as $N \rightarrow \infty$ from earlier estimates and the fact that $\|G_i^M(p^N)\|_N$ is bounded for all N , $1 \leq i \leq l$.

Now using the operator π^N instead of γ^N , we observe that for fixed M ,

$$\begin{aligned}
(3.24) \quad & \|P^M(p^N)\pi^N z - \pi^N P^M(p)z\|_N \leq \|P^M(p^N)(\pi^N - \gamma^N)z\|_N \\
& + \|P^M(p^N)\gamma^N z - \gamma^N P^M(p)z\|_N + \|(\pi^N - \gamma^N)P^M(p)z\|_N,
\end{aligned}$$

where the first term satisfies (again using the orthogonal basis elements for $H^M(p^N)$)

$$\begin{aligned}
\|P^M(p^N)(\pi^N - \gamma^N)z\|_N & \leq k \sum_{i=1}^l |(\pi^N - \gamma^N)z|_N \|G_i^M(p^N)\|_N^2 \\
& + \sum_{i=l+1}^m |(\pi^N - \gamma^N)z|_N \|G_i^M(p^N)\|_N,
\end{aligned}$$

and converges to zero as $N \rightarrow \infty$. For each fixed M , the remaining terms in (3.24) converge to zero as well (using, in the last term, the fact that π^N approximates γ^N in the $V(p^N)$ norm when applied to (fixed) $P^M(p)z$ in $V_0(p)$). Thus the proof of the lemma is complete. \square

We are now able to prove that, for each fixed M , we have

$$\|u^M(t; p^N, q^N) - \pi^N u^M(t; p, q)\|_N \rightarrow 0$$

whenever $(p^N, q^N) \rightarrow (p, q)$. Again, we argue convergence of resolvent operators and then apply the Trotter-Kato theorem to obtain semigroup convergence; owing to the properties of projections given in the last lemma, the calculations we present below differ little from those used above in the proofs of Theorems 3.2 and 3.3.

THEOREM 3.5. *Let $M = 1, 2, \dots$ be fixed, and let $(p^N, q^N) \rightarrow (p, q)$ in $\mathcal{P} \times \mathcal{Q}$. Then for $\lambda = \lambda_0$ and arbitrary $z^M \in H^M(p)$, we have*

$$\|R_\lambda(A^M(p^N, q^N))P^M(p^N)\pi^N z^M - \pi^N R_\lambda(A^M(p, q))z^M\|_N \rightarrow 0,$$

as $N \rightarrow \infty$.

Proof. For fixed M and arbitrary $z \equiv z^M \in H^M(p)$, we define the variables $w = w(p, q, M) \in H^M(p) \subseteq V_0(p)$ and $w^N = w^N(p^N, q^N, M) \in H^M(p^N) \subseteq V_0(p^N)$ as follows:

$$w = R_\lambda(A^M(p, q))z, \quad w^N = R_\lambda(A^M(p^N, q^N))P^M(p^N)\pi^N z.$$

Using the triangle inequality to write

$$\|w^N - \pi^N w\|_N \leq \|w^N - P_V^M(p^N)\gamma^N w\|_N + \|(P_V^M(p^N) - I)\gamma^N w\|_N + \|(\gamma^N - \pi^N)w\|_N,$$

it is clear that, to verify that $\|w^N - \pi^N w\|_N \rightarrow 0$ as $N \rightarrow \infty$, we need only argue that the first term converges, or that $\|w^N - \hat{w}^N\|_N \rightarrow 0$, where here $\hat{w}^N \equiv P_V^M(p^N)\gamma^N w$. Using the notation $z^N = w^N - \hat{w}^N \in H^M(p^N)$, we have for $\lambda = \lambda_0$

$$\begin{aligned} c_1 \|z^N\|_N^2 &\leq \sigma(p^N, q^N)(z^N, z^N) + \lambda |z^N|_N^2 \\ &\leq T_0^N + T_1^N + T_2^N + T_3^N, \end{aligned}$$

where

$$\begin{aligned} T_0^N &\equiv \sigma(p^N, q^N)(w^N, z^N) - \sigma(p, q)(w, P_V^M(p)\gamma(p, p^N)z^N) + \lambda |z^N|_N^2, \\ T_1^N &\equiv |\sigma(p, q)(w, (P_V^M(p) - I)\gamma(p, p^N)z^N)|, \\ T_2^N &\equiv |\sigma(p, q)(w, \gamma(p, p^N)z^N) - \sigma(p^N, q)(\gamma^N w, z^N)|, \\ T_3^N &\equiv |\sigma(p^N, q)(\gamma^N w, z^N) - \sigma(p^N, q^N)(\hat{w}^N, z^N)|. \end{aligned}$$

The terms T_2^N and T_3^N are defined exactly as in the proof of Theorem 3.2, and, in fact, the estimates for each are unchanged from that proof. From the boundedness of $\sigma(p, q)$ and the estimates given for P_V^M given in (3.5), we observe that

$$T_1^N \leq c_2 K_1(M) C_\gamma(p^N, p) \|w\|_p \|z^N\|_N,$$

where $C_\gamma(p^N, p) \rightarrow 0$ as $N \rightarrow \infty$.

It remains to consider T_0^N . We use the fact that w^N and z^N are in $H^M(p^N)$, while w and $P_V^M(p)\gamma(p, p^N)z^N$ are in $H^M(p)$, and write

$$\begin{aligned} T_0^N &= \langle (\lambda I - A^M(p^N, q^N) - \lambda I)w^N, z^N \rangle_N - \langle (\lambda I - A^M(p, q) - \lambda I)w, P_V^M(p)\gamma(p, p^N)z^N \rangle_p \\ &\quad + \lambda |z^N|_N^2 \\ &= \{ \langle P^M(p^N)\pi^N z, z^N \rangle_N - \langle z, P_V^M(p)\gamma(p, p^N)z^N \rangle_p \} \\ &\quad - \lambda \{ \langle w^N, z^N \rangle_N - \langle w, P_V^M(p)\gamma(p, p^N)z^N \rangle_p - |z^N|_N^2 \} \\ &\leq \tau_1^N + \tau_2^N, \end{aligned}$$

where

$$\begin{aligned} \tau_1^N &\equiv \{ \langle P^M(p^N)\pi^N z, z^N \rangle_N - \langle z, \gamma(p, p^N)z^N \rangle_p \} \\ &\quad - \lambda \{ \langle w^N, z^N \rangle_N - \langle w, \gamma(p, p^N)z^N \rangle_p - |z^N|_N^2 \}, \\ \tau_2^N &\equiv |\langle z, (P_V^M(p) - I)\gamma(p, p^N)z^N \rangle_p| + |\lambda| |\langle w, (P_V^M(p) - I)\gamma(p, p^N)z^N \rangle_p|. \end{aligned}$$

If we define $P^N \equiv P^M(p^N)$ (no confusion should result because M is fixed), we observe that τ_1^N is identical to T_1^N as given by (3.12) in the proof of Theorem 3.2, and that τ_1^N may be bounded as T_1^N is in that theorem. For τ_2^N , we again use the properties of P_V^M and find that

$$\tau_2^N \leq (|z|_p + |\lambda| \|w\|_p) k K_1(M) C_\gamma(p^N, p) \|z^N\|_N.$$

Combining the above estimates (and using arguments like those in Theorem 3.2), we find $\|z^N\|_N \rightarrow 0$ as $N \rightarrow \infty$, or that the statement of the theorem is valid. \square

At this point, it is not difficult to argue, using the Trotter–Kato theorem that $(p^N, q^N) \rightarrow (p, q) \in \mathcal{P} \times \mathcal{Q}$ implies, for fixed M , $|u^M(t; p^N, q^N) - \pi^N u^M(t; p, q)|_N \rightarrow 0$ as $N \rightarrow \infty$, at a rate uniform in $t \geq 0$ in compact subintervals. In fact, defining $X \equiv (H^M(p), |\cdot|_p)$, $X^N \equiv (H^M(p^N), |\cdot|_N)$, $A \equiv A^M(p, q)$, $A^N \equiv A^M(p^N, q^N)$, $B^N \equiv P^N \pi^N$, and $P^N \equiv P^M(p^N)$, we observe that (i) and (ii) of the Trotter–Kato theorem hold, while (iii) follows immediately from the triangle inequality and previous estimates. We thus have

$$|T^M(t; p^N, q^N) P^M(p^N) \pi^N z^M - P^M(p^N) \pi^N T^M(t; p, q) z^M|_N \rightarrow 0 \quad \text{as } N \rightarrow \infty,$$

uniformly in $t \in [0, T]$, for fixed M and arbitrary $z^M \in H^M(p)$, so that

$$|T^M(t; p^N, q^N) P^M(p^N) \pi^N z^M - \pi^N T^M(t; p, q) z^M|_N \rightarrow 0 \quad \text{as } N \rightarrow \infty.$$

Finally, to argue convergence of $|u^M(t; p^N, q^N) - \pi^N u^M(t; p, q)|_N$, we consider the convergence of the initial condition term in the semigroup representation for $u^M(t; p^N, q^N) - \pi^N u^M(t; p, q)$ (estimates for the integral terms are similar). This term satisfies

$$\begin{aligned} (3.25) \quad & |T^M(t; p^N, q^N) P^M(p^N) \pi^N u_0 - \pi^N T^M(t; p, q) P^M(p) u_0|_N \\ & \leq |T^M(t; p^N, q^N) (P^M(p^N) \pi^N u_0 - \pi^N P^M(p) u_0)|_N \\ & \quad + |T^M(t; p^N, q^N) (I - P^M(p^N)) \pi^N (P^M(p) u_0)|_N \\ & \quad + |T^M(t; p^N, q^N) P^M(p^N) \pi^N (P^M(p) u_0) - \pi^N T^M(t; p, q) (P^M(p) u_0)|_N, \end{aligned}$$

where it is easily seen that each term in this expression converges to zero as $N \rightarrow \infty$, using the boundedness of $T^M(t; p^N, q^N)$ and the properties of projections ((3.23) and (3.8), with argument $P^M(p) u_0 \in H^M(p)$ in the second term above). Repeating these arguments for the integral terms, we obtain the convergence

$$|u^M(t; p^N, q^N) - \pi^N u^M(t; p, q)|_N \rightarrow 0 \quad \text{as } N \rightarrow \infty$$

for each fixed M .

Again, however, it is our goal to obtain a corresponding statement of continuous dependence in the stronger V topology; this finding is stated in the theorem that follows.

THEOREM 3.6. *Let M be fixed and let $(p^N, q^N) \rightarrow (p, q) \in \mathcal{P} \times \mathcal{Q}$ as $N \rightarrow \infty$. Then, for arbitrary $z^M \in H^M(p)$,*

$$\|T^M(t; p^N, q^N) P^M(p^N) \pi^N z^M - \pi^N T^M(t; p, q) z^M\|_N \rightarrow 0 \quad \text{as } N \rightarrow \infty,$$

uniformly in compact subintervals of $t > 0$. Furthermore, we have, for each fixed M ,

$$\|u^M(t; p^N, q^N) - \pi^N u^M(t; p, q)\|_N \rightarrow 0 \quad \text{as } N \rightarrow \infty,$$

at a rate uniform in $t > 0$ in compact subintervals.

Proof. The proof changes little from the proof of Theorem 3.3; we sketch only the fundamental differences here. Once again, we appeal to the Trotter–Kato theorem, employing now $X \equiv (H^M(p), \|\cdot\|_p)$, $X^N \equiv (H^M(p^N), \|\cdot\|_N)$, $A \equiv A^M(p, q)$, $A^N \equiv A^M(p^N, q^N)$, and $B^N \equiv P_V^N \pi^N$ (where we have also abbreviated $P_V^N \equiv P_V^M(p^N)$).

The proofs of (i) and (ii) in the Trotter–Kato theorem differ little from the proofs of the analogues of these steps in Theorem 3.3; we note that for (iii) we may also rely on the calculations in the proof of that theorem (specifically, (3.21)), if we use the fact that

$$|(P_V^M(p^N) - P^M(p^N))\pi^N z^M|_N \rightarrow 0 \quad \text{as } N \rightarrow \infty$$

and that

$$\|(I - P_V^N)\pi^N R_\lambda(A)z^M\|_N \rightarrow 0 \quad \text{as } N \rightarrow \infty$$

(using $R_\lambda(A)z^M \in H^M(p) \subseteq V_0(p)$); thus we are able to claim that (iii) of the Trotter–Kato theorem indeed holds. It follows then that for fixed M and arbitrary $z^M \in H^M(p)$,

$$\|T^M(t; p^N, q^N)P_V^M(p^N)\pi^N z^M - P_V^M(p^N)\pi^N T^M(t; p, q)z^M\|_N \rightarrow 0 \quad \text{as } N \rightarrow \infty,$$

and that a similar expression involving the $H(p^N)$ projection operator $P^M(p^N)$ then satisfies

$$\begin{aligned} & \|T^M(t; p^N, q^N)P^M(p^N)\pi^N z^M - \pi^N T^M(t; p, q)z^M\|_N \\ & \leq \|T^M(t; p^N, q^N)(P^M(p^N) - P_V^M(p^N))\pi^N z^M\|_N \\ & \quad + \|T^M(t; p^N, q^N)P_V^M(p^N)\pi^N z^M - P_V^M(p^N)\pi^N T^M(t; p, q)z^M\|_N \\ & \quad + \|(P_V^M(p^N) - I)\pi^N T^M(t; p, q)z^M\|_N. \end{aligned}$$

Because the first term on the right satisfies

$$\begin{aligned} & \|T^M(t; p^N, q^N)(P^M(p^N) - P_V^M(p^N))\pi^N z^M\|_N \leq \\ & \leq \tilde{\mathcal{M}}e^{\lambda_0 t} t^{-1/2} |(P^M(p^N) - P_V^M(p^N))\pi^N z^M|_N \end{aligned}$$

(using estimates from the proof of Theorem 3.3), it is clear then that all terms above converge to zero as $N \rightarrow \infty$, uniformly in $t > 0$ in compact subintervals. Finally, repeating arguments similar to those given above, and like those used to argue $|u^M(t; p^N, q^N) - \pi^N u^M(t; p, q)|_N \rightarrow 0$ as $N \rightarrow \infty$ (here using the estimate

$$\begin{aligned} & \|T^M(t; p^N, q^N)(I - P^M(p^N))\pi^N(P^M(p)u_0)\|_N \\ & \leq \tilde{\mathcal{M}}e^{\lambda_0 t} t^{-1/2} |(I - P^M(p^N))\pi^N(P^M(p)u_0)|_N \end{aligned}$$

in (3.25)), we conclude that the remainder of the theorem is valid. \square

4. A generalized approximation framework for second-order problems.

We consider here an application of the parameter estimation ideas developed in previous sections to second-order systems of the form discussed in [3], modified as before for parameter-dependent state spaces. Specifically, we investigate the problem of estimating parameters p and q , which appear in the abstract equation in $H(p)$,

$$(4.1) \quad \ddot{u}(t) + B(p, q)\dot{u}(t) + A(p, q)u(t) = F(t; p, q),$$

where appropriate initial conditions are given and, as in [3], the operators $A(p, q)$ and $B(p, q)$ are defined through parameter-dependent sesquilinear forms that satisfy properties to be specified below.

The modifications that we make to the arguments found in [3] for the second-order case are entirely analogous to those used to develop the first-order theory in previous sections; for this reason, we only outline the basic results of Banks and Ito, indicating the changes required for the parameter-dependent state space problem.

Throughout this section, we assume that Hypotheses 1 and 3 (from §3) hold, and that for every $(p, q) \in \mathcal{P} \times \mathcal{Q}$, the sesquilinear form $\sigma_1(p, q)$ is symmetric and satisfies assumptions (a)–(c) of Hypothesis 2, with $\lambda_0 = 0$ in 2(c). The sesquilinear form $\sigma_2(p, q)$, which need not be symmetric, is assumed to satisfy (a) and (b) of Hypothesis 2, and a weaker form of (c) of the same hypothesis, given below.

(c') *Uniform (in (p, q)) $H(p)$ -semicoercivity*: There exists $b \geq 0$ such that for every $(p, q) \in \mathcal{P} \times \mathcal{Q}$ and all $v_0 \in V_0(p)$,

$$\operatorname{Re} \sigma_2(p, q)(v_0, v_0) \geq b|v_0|_p^2.$$

As in [3], for each $(p, q) \in \mathcal{P} \times \mathcal{Q}$, the sesquilinear forms $\sigma_1(p, q)$ and $\sigma_2(p, q)$ may be used to define continuous linear operators $A(p, q)$ and $B(p, q)$, respectively, each mapping $V_0(p)$ to $V_0(p)^*$, satisfying

$$\begin{aligned} \sigma_1(p, q)(u, v) &= (A(p, q)u, v)_{V_0(p)^*, V_0(p)}, \\ \sigma_2(p, q)(u, v) &= (B(p, q)u, v)_{V_0(p)^*, V_0(p)}, \end{aligned}$$

for $u, v \in V_0(p)$. (Here $(\cdot, \cdot)_{V_0(p)^*, V_0(p)}$ denotes the duality pairing, which is the unique continuous extension of the $H(p)$ inner product from $H(p) \times V_0(p)$ to $V_0(p)^* \times V_0(p)$.) As discussed in [3], the operator $A(p, q)$ is associated with the usual “stiffness” operator found in applications, while $B(p, q)$ corresponds to a “damping” operator; additionally, both operators may be viewed as densely defined operators in $H(p)$.

We define $\mathcal{V}_0(p) \equiv V_0(p) \times V_0(p)$, $\mathcal{H}_0(p) \equiv V_0(p) \times H(p)$, $\mathcal{V}(p) \equiv V(p) \times V(p)$, and $\mathcal{H}(p) \equiv V(p) \times H(p)$, and rewrite (4.1) in first-order form in the variable $w(t) = (u(t), v(t)) \in \mathcal{V}_0(p)$; in a variational framework, this equation may be expressed as follows:

$$(4.2) \quad \langle \dot{w}(t), \chi \rangle_{\mathcal{H}(p)} + \sigma(p, q)(w(t), \chi) = \langle F(t), \chi \rangle_{\mathcal{H}(p)}, \quad \chi \in \mathcal{V}_0(p),$$

where we now use $\sigma(p, q)$ to denote the sesquilinear form, defined on the product $\mathcal{V}_0(p) \times \mathcal{V}_0(p)$, which is given by

$$(4.3) \quad \sigma(p, q)((u, v), (\phi, \psi)) = -(v, \phi)_{V(p)} + \sigma_1(p, q)(u, \psi) + \sigma_2(p, q)(v, \psi).$$

Alternatively, (4.1) may be written in operator form as

$$(4.4) \quad \dot{w}(t) = \mathcal{A}(p, q)w(t) + \mathcal{F}(t; p, q),$$

where $\mathcal{F}(t; p, q) = (0, F(t; p, q))$ and $\mathcal{A}(p, q)$ is the usual operator defined from $\sigma(p, q)$; we note that [3] the operator $\mathcal{A}(p, q)$ may be expressed in $\mathcal{H}_0(p)$ as

$$\mathcal{A}(p, q) = \begin{bmatrix} 0 & I \\ -A(p, q) & -B(p, q) \end{bmatrix},$$

with $\operatorname{dom} \mathcal{A}(p, q) = \{(\phi, \psi) \in \mathcal{H}_0(p) | \psi \in V_0(p), A(p, q)\phi + B(p, q)\psi \in H(p)\}$, $\operatorname{dom} \mathcal{A}(p, q) \subseteq \mathcal{V}_0(p)$. Using arguments similar to those found in [3], it follows that $\mathcal{A}(p, q)$ is the infinitesimal generator of a C_0 -semigroup $\mathcal{T}(t; p, q)$ on $\mathcal{H}_0(p)$; it is this semigroup for

which a parameter estimation theory, similar to that developed in the previous sections, will be formulated. In keeping with [3], we do not directly apply the theory already developed for first-order problems to this problem because the assumption of coercivity that was made on the underlying sesquilinear form in first-order problems would necessitate that *both* $\sigma_1(p, q)$ and $\sigma_2(p, q)$ be V_0 -coercive; indeed, an attractive feature of the work found in [3] is that the weaker *semicoercivity* assumption for $\sigma_2(p, q)$ may be used, allowing for very general types of damping terms that are often desired in applications. We take the same approach here.

Approximation and convergence arguments will again be based on an application of the Trotter–Kato theorem. To this end, we will need some operators that are second-order analogues of those used in earlier sections; we define the operator $\mathcal{P}^N(p)$ as the orthogonal projection of $\mathcal{H}(p)$ onto $\mathcal{H}^N(p)$, and the space-changing maps $\Pi(\tilde{p}, p) = \pi(\tilde{p}, p) \times \pi(\tilde{p}, p)$ and $\Gamma(\tilde{p}, p) = \gamma(\tilde{p}, p) \times \gamma(\tilde{p}, p)$. As usual, we let $\mathcal{A}^N(p, q)$ denote the operator defined by the restriction of $\sigma(p, q)$ to $\mathcal{H}^N(p)$, where $\mathcal{H}^N(p)$ is the (quotient) approximation space defined by $\mathcal{H}^N(p) = H^N(p) \times H^N(p)$, $\mathcal{H}^N(p) \subseteq \mathcal{V}_0(p)$. As in [3], our first result is a statement of resolvent convergence in a norm that is stronger than that required for a later application of the Trotter–Kato theorem.

THEOREM 4.1. *Assume that the above conditions hold, and let $(p^N, q^N) \rightarrow (p, q)$ in $\mathcal{P} \times \mathcal{Q}$. Then for $\lambda > 0$ and arbitrary $\xi \in \mathcal{H}_0(p)$, we have*

$$|R_\lambda(\mathcal{A}^N(p^N, q^N))\mathcal{P}^N(p^N)\Pi(p^N, p)\xi - \Pi(p^N, p)R_\lambda(\mathcal{A}(p, q))\xi|_{\mathcal{V}(p^N)} \rightarrow 0,$$

as $N \rightarrow \infty$.

Proof. Let $\lambda > 0$ and $\xi = (\eta, \nu)$ be arbitrarily given in $\mathcal{H}_0(p)$. The existence, for $\lambda > 0$, of the resolvent operator will be demonstrated below, so there is no difficulty in defining $w = w(p, q) \in \mathcal{V}_0(p)$ and $w^N = w^N(p^N, q^N) \in \mathcal{H}^N(p^N) \subseteq \mathcal{V}_0(p^N)$ through the relations

$$\begin{aligned} w &\equiv (\phi, \psi) = R_\lambda(\mathcal{A})\xi, \\ w^N &\equiv (\phi^N, \psi^N) = R_\lambda(\mathcal{A}^N)\mathcal{P}^N\Pi^N\xi, \end{aligned}$$

where here we have made the usual notational abbreviations. It follows from the definition of $\mathcal{A}(p, q)$ that (ϕ, ψ) satisfies

$$(4.5) \quad \lambda\phi - \psi = \eta,$$

$$(4.6) \quad \lambda\psi + A(p, q)\phi + B(p, q)\psi = \nu,$$

while (ϕ^N, ψ^N) satisfies

$$(4.7) \quad \lambda\phi^N - \psi^N = \eta^N,$$

$$(4.8) \quad \lambda\psi^N + A^N(p, q)\phi^N + B^N(p, q)\psi^N = \nu^N,$$

where $(\eta^N, \nu^N) \equiv \mathcal{P}^N\Pi^N(\eta, \nu)$.

We consider (4.5) and (4.6) in (ϕ, ψ) : substituting the value of ψ given in the first equation into the second equation, we find that remaining is an equation (in $(V_0(p))^*$) in the unknown ϕ ,

$$(4.9) \quad \lambda^2\phi + A(p, q)\phi + \lambda B(p, q)\phi = \nu + \lambda\eta + B(p, q)\eta,$$

which is equivalent [3] to the following equation:

$$(4.10) \quad \sigma_\lambda(p, q)(\phi, \zeta) = \langle \nu, \zeta \rangle_p + \lambda \langle \eta, \zeta \rangle_p + \sigma_2(p, q)(\eta, \zeta), \quad \zeta \in V_0(p);$$

here the sesquilinear form $\sigma_\lambda(p, q) : V_0(p) \times V_0(p) \rightarrow \mathcal{C}$ is defined as it is in [3], namely,

$$\sigma_\lambda(p, q)(\phi, \zeta) \equiv \lambda^2 \langle \phi, \zeta \rangle_p + \sigma_1(p, q)(\phi, \zeta) + \lambda \sigma_2(p, q)(\phi, \zeta).$$

It is easily seen that for each $\lambda > 0$, the form σ_λ satisfies assumptions (a) and (b) of Hypothesis 2 (continuity in parameters, uniform $V_0(p)$ -boundedness) given in the last section; additionally, arguments similar to those in [3] may be used to establish the following coercivity estimate:

$$\sigma_\lambda(p, q)(\phi, \phi) > c_1 \|\phi\|_p^2, \quad \phi \in V_0(p), \lambda > 0,$$

which holds uniformly in $(p, q) \in \mathcal{P} \times \mathcal{Q}$. It thus follows [3] that, for each $\lambda > 0$, (4.10) is solvable for $\phi \in V_0(p)$, that $w = (\phi, \psi)$ (with $\psi = \lambda\phi - \eta$) belongs in $\text{dom } \mathcal{A}(p, q)$ for arbitrary $(\eta, \nu) \in \mathcal{H}_0(p)$, and that $R_\lambda(\mathcal{A})$ exists as an element in $\mathcal{L}(\mathcal{H}_0(p))$.

Similar arguments may be made to show that (ϕ^N, ψ^N) satisfies

$$(4.11) \quad \lambda\phi^N - \psi^N = \eta^N,$$

$$(4.12) \quad \sigma_\lambda(p^N, q^N)(\phi^N, \zeta^N) = \langle \nu^N, \zeta^N \rangle_N + \lambda \langle \eta^N, \zeta^N \rangle_N + \sigma_2(p^N, q^N)(\eta^N, \zeta^N)$$

for all $\zeta^N \in \mathcal{H}^N(p^N)$.

As a first step in obtaining the results of the theorem, i.e., that $|\Pi^N w - w^N|_{\mathcal{V}(p^N)} \rightarrow 0$, we will demonstrate that $\|\pi^N \phi - \phi^N\|_N \rightarrow 0$ as $N \rightarrow \infty$; using standard arguments it is easily seen that it suffices to show that $\|\phi^N - \hat{\phi}^N\|_N \rightarrow 0$, where $\hat{\phi}^N \equiv P_V^N \gamma^N \phi$.

Let $\zeta^N = \phi^N - \hat{\phi}^N \in H^N(p^N)$. Coercivity of σ_λ on $V_0(p^N)$ and an application of the triangle inequality yield the following estimates:

$$\begin{aligned} c_1 \|\zeta^N\|_N^2 &< \sigma_\lambda(p^N, q^N)(\zeta^N, \zeta^N) \\ &= \sigma_\lambda(p^N, q^N)(\phi^N, \zeta^N) - \sigma_\lambda(p^N, q^N)(\hat{\phi}^N, \zeta^N) \\ &\leq T_1^N + T_2^N + T_3^N + T_4^N, \end{aligned}$$

where

$$\begin{aligned} T_1^N &\equiv |\sigma_\lambda(p^N, q^N)(\phi^N, \zeta^N) - \sigma_\lambda(p, q)(\phi, \gamma(p, p^N)\zeta^N)|, \\ T_2^N &\equiv |\sigma_\lambda(p, q)(\phi, \gamma(p, p^N)\zeta^N) - \sigma_\lambda(p^N, q)(\gamma^N \phi, \zeta^N)|, \\ T_3^N &\equiv |\sigma_\lambda(p^N, q)(\gamma^N \phi, \zeta^N) - \sigma_\lambda(p^N, q^N)(\gamma^N \phi, \zeta^N)|, \\ T_4^N &\equiv |\sigma_\lambda(p^N, q^N)(\gamma^N \phi - \hat{\phi}^N, \zeta^N)|. \end{aligned}$$

To estimate the first term we may use (4.10) and (4.12) to write

$$\begin{aligned} T_1^N &\leq |\langle \nu^N + \lambda\eta^N, \zeta^N \rangle_N - \langle \nu + \lambda\eta, \gamma(p, p^N)\zeta^N \rangle_p| \\ &\quad + |\sigma_2(p^N, q^N)(\eta^N, \zeta^N) - \sigma_2(p, q)(\eta, \gamma(p, p^N)\zeta^N)| \\ &\leq |\langle (\nu^N + \lambda\eta^N) - \gamma^N(\nu + \lambda\eta), \zeta^N \rangle_N| + |\langle \gamma^N(\nu + \lambda\eta), \zeta^N \rangle_N - \langle \nu + \lambda\eta, \gamma(p, p^N)\zeta^N \rangle_p| \\ &\quad + |\sigma_2(p^N, q^N)(\eta^N - \gamma^N \eta, \zeta^N)| + |\sigma_2(p^N, q^N)(\gamma^N \eta, \zeta^N) - \sigma_2(p^N, q)(\gamma^N \eta, \zeta^N)| \\ &\quad + |\sigma_2(p^N, q)(\gamma^N \eta, \zeta^N) - \sigma_2(p, q)(\eta, \gamma(p, p^N)\zeta^N)| \\ &\leq C \|\zeta^N\|_N (|\nu^N - \gamma^N \nu|_N + \|\eta^N - \gamma^N \eta\|_N + C_p(p^N) + C_q(q^N) + C_{p,q}(p^N)), \end{aligned}$$

where $C > 0$ is a constant and we have used the continuity properties of both the sesquilinear form σ_2 and the $H(p)$ inner product in the last inequality. Furthermore, we note that two terms in this bound may be estimated using $\|\eta^N - \gamma^N \eta\|_N^2 + |\nu^N - \gamma^N \nu|_N^2 = |\mathcal{P}^N \Pi^N(\eta, \nu) - \Gamma^N(\eta, \nu)|_{\mathcal{H}(p^N)}^2 \leq 2|(\mathcal{P}^N - I)\Pi^N(\eta, \nu)|_{\mathcal{H}(p^N)}^2 +$

$$2|(\Pi^N - \Gamma^N)(\eta, \nu)|_{\mathcal{H}(p^N)}^2 \rightarrow 0 \quad \text{as } N \rightarrow \infty.$$

Finally, using the properties of continuity and boundedness for σ_λ , we find that similar calculations may be made for T_2^N , T_3^N , and T_4^N ; we may put these results together to obtain $c_1 \|\zeta^N\|_N^2 \leq k(N) \|\zeta^N\|_N$, for some real scalar $k(N)$ that satisfies $k(N) \rightarrow 0$ as $N \rightarrow \infty$. Therefore, $\|\zeta^N\|_N \rightarrow 0$ as $N \rightarrow \infty$.

It thus follows that $\|\pi^N \phi - \phi^N\|_N \rightarrow 0$. Appealing to (4.5) and (4.7), we also have

$$\begin{aligned} \|\pi^N \psi - \psi^N\|_N &\leq |\lambda| \|\pi^N \phi - \phi^N\|_N + \|\pi^N \eta - \eta^N\|_N \\ &\leq |\lambda| \|\pi^N \phi - \phi^N\|_N + \|(\pi^N - \gamma^N) \eta\|_N + \|\gamma^N \eta - \eta^N\|_N, \end{aligned}$$

so that we may conclude $|\Pi^N w - w^N|_{\mathcal{V}(p^N)}^2 = \|\pi^N \phi - \phi^N\|_N^2 + \|\pi^N \psi - \psi^N\|_N^2 \rightarrow 0$ as $N \rightarrow \infty$. \square

We let $\mathcal{T}^N(t; p^N, q^N)$ denote the C_0 -semigroup in $\mathcal{H}(p^N)$, which is generated by $\mathcal{A}^N(p^N, q^N)$, and apply the Trotter-Kato theorem with $X = \mathcal{H}_0(p)$, $X^N = \mathcal{H}(p^N)$, and $B^N = \mathcal{P}^N \Pi^N$. It is easily verified that the following statement of semigroup convergence is then obtained (which is stated using our conventional notational shorthand).

THEOREM 4.2. *For arbitrary $\xi \in \mathcal{H}_0(p)$,*

$$|\mathcal{T}^N(t) \mathcal{P}^N \Pi^N \xi - \Pi^N \mathcal{T}(t) \xi|_{\mathcal{H}(p^N)} \rightarrow 0,$$

as $N \rightarrow \infty$, uniformly in $t \in [0, T]$.

We thus have the results needed to claim (under appropriate assumptions on initial conditions and external force $F(t; p, q)$) that the “state variable convergence” step (step 2 in §2.2) has been demonstrated (i.e., $(p^N, q^N) \rightarrow (p, q)$ implies that $\|u^N(t; p^N, q^N) - \pi^N u(t; p, q)\|_N \rightarrow 0$ and $\|v^N(t; p^N, q^N) - \pi^N v(t; p, q)\|_N \rightarrow 0$, uniformly in $t \in [0, T]$). The remaining continuous dependence result (step 1 of that section) is also easily established, using techniques similar to those employed above. For completeness, we provide a statement of these findings, without proof, in the theorem that concludes this section.

THEOREM 4.3. *Let $M = 1, 2, \dots$ be fixed and let $(p^N, q^N) \rightarrow (p, q)$ in $\mathcal{P} \times \mathcal{Q}$. Then for $\lambda > 0$ and arbitrary $\xi^M \in \mathcal{H}^M(p)$, we have*

$$|R_\lambda(\mathcal{A}^M(p^N, q^N)) \mathcal{P}^M(p^N) \Pi^N \xi^M - \Pi^N R_\lambda(\mathcal{A}^M(p, q)) \xi^M|_{\mathcal{V}(p^N)} \rightarrow 0,$$

as $N \rightarrow \infty$. It further follows that, for each fixed M ,

$$|\mathcal{T}^M(t; p^N, q^N) \mathcal{P}^M(p^N) \Pi^N \xi^M - \Pi^N \mathcal{T}^M(t; p, q) \xi^M|_{\mathcal{H}(p^N)} \rightarrow 0,$$

as $N \rightarrow \infty$, uniformly in $t \in [0, T]$, so that,

$$\begin{aligned} \|u^M(t; p^N, q^N) - \pi^N u^M(t; p, q)\|_N &\rightarrow 0, \\ \|v^M(t; p^N, q^N) - \pi^N v^M(t; p, q)\|_N &\rightarrow 0, \end{aligned}$$

as $N \rightarrow \infty$, uniformly in $t \in [0, T]$.

5. Application to domain shape estimation problem. We verify here that the assumptions made in §3 and in Remark 2.3 are satisfied in the case of the domain shape estimation problem defined in §2. We will use operators and spaces exactly as given there.

Before turning to the hypotheses in §3, we justify those conditions (as stated in Remark 2.3) needed if J_t is to be used as the data-fitting criterion. That is, we must verify that $\pi(p, p) = I$ (where I is the identity map), and that the map $p \rightarrow \|\pi_t(p)v\|_p : \mathcal{P} \rightarrow \mathbb{R}$ is continuous for any $v \in H^1(\Omega)$. Indeed, while the former is trivial ($\pi(p, p) \equiv \pi_t(p)\pi_e(p) = I$), for the latter we need only observe that, for $p, \tilde{p} \in \mathcal{P}$,

$$\begin{aligned}
 (5.1) \quad & \left| \|\pi_t(p)v\|_p^2 - \|\pi_t(\tilde{p})v\|_{\tilde{p}}^2 \right| \leq \left| \int_{\Omega_p} |Dv|^2 - \int_{\Omega_{\tilde{p}}} |Dv|^2 \right| + \left| \int_{\Omega_p} |v|^2 - \int_{\Omega_{\tilde{p}}} |v|^2 \right| \\
 & = \left| \int_{\Omega_p \setminus (\Omega_p \cap \Omega_{\tilde{p}})} |Dv|^2 - \int_{\Omega_{\tilde{p}} \setminus (\Omega_p \cap \Omega_{\tilde{p}})} |Dv|^2 \right| \\
 & \quad + \left| \int_{\Omega_p \setminus (\Omega_p \cap \Omega_{\tilde{p}})} |v|^2 - \int_{\Omega_{\tilde{p}} \setminus (\Omega_p \cap \Omega_{\tilde{p}})} |v|^2 \right| \\
 & \leq 2 \int_{\mathcal{R}(\Omega_p, \Omega_{\tilde{p}})} |Dv|^2 + 2 \int_{\mathcal{R}(\Omega_p, \Omega_{\tilde{p}})} |v|^2 \\
 & < \epsilon,
 \end{aligned}$$

provided that $d_p(\tilde{p}, p) < \delta = \delta(\epsilon, v)$ (where $\mathcal{R}(\Omega_p, \Omega_{\tilde{p}}$ was defined in §2). Thus, all assumptions stated in Remark 2.3 are seen to hold.

For $H(p) \equiv L_2(\Omega_p)$, $V(p) \equiv H^1(\Omega_p)$, and $V_0(p) \equiv H_0^1(\Omega_p)$ (or $H^1(\Omega_p)$), we have that $V_0(p) \subset V(p) \subset H(p)$, $V_0(p)$ dense and continuously imbedded in $H(p)$, and $V_0(p)$ a closed subspace of $V(p)$. In addition, $|v|_p \leq \|v\|_p$ so that Hypothesis 1(a) holds.

From the way that the maps $\gamma_t(p)$, $\gamma_e(p)$ were defined in §2.2, we see that these maps are isomorphisms, with $\gamma_e(p) \equiv \gamma_t(p)^{-1}$, and

$$\begin{aligned}
 c_1|u| &\leq |\gamma_t(p)u|_p \leq C_1|u|, & u \in H, \\
 c_2\|u\| &\leq \|\gamma_t(p)u\|_p \leq C_2\|u\|, & u \in V,
 \end{aligned}$$

where positive constants c_1 , C_1 , c_2 , C_2 are independent of p , u (see [1, pp. 63–64]). It follows easily then that all conditions in Hypothesis 1(b) are satisfied.

To verify Hypothesis 1(c), we observe that, for fixed $p \in \mathcal{P}$ and any $v_0 \in V_0(p)$,

$$\begin{aligned}
 \|(\pi(\tilde{p}, p) - \gamma(\tilde{p}, p))v_0\|_{H^1(\Omega_{\tilde{p}})} &= \|(I - \gamma(\tilde{p}, p))v_0\|_{H^1(\Omega_p \cap \Omega_{\tilde{p}})} \\
 &\quad + \|(\pi(\tilde{p}, p) - \gamma(\tilde{p}, p))v_0\|_{H^1(\Omega_{\tilde{p}} \setminus (\Omega_p \cap \Omega_{\tilde{p}}))},
 \end{aligned}$$

where the second term may be made arbitrarily small whenever $d_p(p, \tilde{p})$ is sufficiently small (the arguments use the uniform, in p, \tilde{p} , boundedness of composite maps π and γ and estimates similar to those found in (5.1) above). For the first term, we may assume (without loss of generality) that $v_0 \in C^\infty(\Omega_p)$. Letting $k(\tilde{p}, p)(x) \equiv \phi(p)(\psi(\tilde{p})(x))$ for $x \in \Omega_p \cap \Omega_{\tilde{p}}$, we have

$$\begin{aligned}
 \|v_0 - \gamma(\tilde{p}, p)v_0\|_{H^1(\Omega_p \cap \Omega_{\tilde{p}})}^2 &= \int_{\Omega_p \cap \Omega_{\tilde{p}}} |v_0(x) - v_0(k(\tilde{p}, p)(x))|^2 \\
 &\quad + \int_{\Omega_p \cap \Omega_{\tilde{p}}} |Dv_0(x) - Dv_0(k(\tilde{p}, p)(x))|^2,
 \end{aligned}$$

where

$$\begin{aligned} Dv_0(x) &= (\partial_1 v_0(x), \dots, \partial_n v_0(x))^T, \\ Dv_0(k(\tilde{p}, p)(x)) &= (\partial_1 v_0(k(\tilde{p}, p)(x)), \dots, \partial_n v_0(k(\tilde{p}, p)(x))) J(\tilde{p}, p)(x), \end{aligned}$$

and we have used $J(\tilde{p}, p)(x)$ to denote the Jacobian matrix associated with $k(\tilde{p}, p)(x)$. It follows from the assumed continuity of the maps $p \rightarrow \phi_i(p)$, $p \rightarrow \psi_i(p)$ (in the prescribed sense), that whenever $d_p(\tilde{p}, p) < \delta(\epsilon, p, v_0)$, we have

$$\begin{aligned} |x - k(\tilde{p}, p)(x)| &< \epsilon, \\ |v_0(x) - v_0(k(\tilde{p}, p)(x))| &< \epsilon, \\ |\partial_i v_0(x) - \partial_i v_0(k(\tilde{p}, p)(x))| &< \epsilon, \quad i = 1, \dots, n, \\ \|J(\tilde{p}, p)(x) - I\| &< \epsilon, \end{aligned}$$

(the Euclidean matrix norm is used in the last inequality) where these estimates are uniform in $x \in \Omega_p \cap \Omega_{\tilde{p}}$. It follows immediately from these inequalities that $\|v_0 - \gamma(\tilde{p}, p)v_0\|_{H^1(\Omega_p \cap \Omega_{\tilde{p}})}$ may be made arbitrarily small whenever \tilde{p} is sufficiently close to p , from which (c) of Hypothesis 1 follows.

Hypothesis 1(d) is easily established using the observation that, for arbitrary $u, v \in L_2(\Omega)$,

$$\begin{aligned} & \left| \int_{\Omega_{\tilde{p}}} \gamma_t(\tilde{p})u \gamma_t(\tilde{p})v - \int_{\Omega_p} \gamma_t(p)u \gamma_t(p)v \right| \\ &= \left| \int_{\Omega_{\tilde{p}}} u(\psi(\tilde{p})(y)) v(\psi(\tilde{p})(y)) dy - \int_{\Omega_p} u(\psi(p)(y)) v(\psi(p)(y)) dy \right| \\ &= \left| \int_{\Omega} u(x)v(x) (\det(\phi(\tilde{p}))'(x) - \det(\phi(p))'(x)) dx \right|; \end{aligned}$$

using the continuity assumed for coordinate transformations, it is easily argued that the last expression may be made as small as desired by taking the distance $d_p(\tilde{p}, p)$ sufficiently small. Thus Hypothesis 1(d) is obtained.

Now that the basic assumptions about spaces and mappings have been verified, the resulting conditions are easily verified. In particular, (a) of Hypothesis 2 (continuity with respect to parameters of the sesquilinear form defined in (2.7)) may be established using steps similar to those used above for (c) of Hypothesis 1; it is a standard result [25] that all other needed properties of the sesquilinear form (uniform coercivity and boundedness) hold for this example.

The construction of approximating spaces $H^N(p)$ is quite simple for this example, given that the theoretical space-changing map $\gamma_t(p)$ is an isomorphism on all of H . To this end, we define for each $N = 1, 2, \dots$, a (fixed) finite-dimensional space $H^N \subseteq V_0$, which has the property that any element $v_0 \in V_0$ may be approximated by $\hat{v}_0^N \in H^N$ for N sufficiently large; i.e., there exists $\hat{v}_0^N \in H^N$ such that

$$\|v_0 - \hat{v}_0^N\|_{H^1(\Omega)} \rightarrow 0 \quad \text{as } N \rightarrow \infty.$$

To cite a typical example, if the original *fixed* domain Ω is a polygonal domain in \mathbb{R}^n , we may use any of a number of “simplicial” or “rectangular” finite element spaces (see, for instance, Chapter 3 of [9]) to construct H^N ; in the case of Dirichlet boundary conditions, we also require that elements in this space are zero on the boundary of Ω (an easily implemented condition for spaces constructed using finite elements).

The parameter-dependent approximation spaces $H^N(p)$ are then defined from the fixed approximation spaces H^N ; i.e., for each $p \in \mathcal{P}$, $H^N(p) \equiv \gamma_t(p)H^N$. It is easily seen that $H^N(p)$ satisfies all needed conditions ((a), (b) of Hypothesis 3) that are required to complete the construction of our approximation theory. We note that if finite elements were used in the construction of H^N , then $H^N(p)$ becomes a finite element space of "irregularly shaped" elements that conform to the boundary of Ω_p .

With all assumptions thus satisfied, we have in place an approximation scheme for the problem of estimating an unknown domain in \mathbb{R}^n . The actual implementation requires an iterative procedure for searching over the parameter space \mathcal{P} (or a discretized version of \mathcal{P}), where at each step in the iteration an evaluation of the approximating equations (and approximating cost functional) is performed. For actual implementation of such algorithms, refer to [17], [20] for one-dimensional examples, while for two-dimensional applications, see [4], [18].

REFERENCES

- [1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
- [2] H. T. BANKS, *On a variational approach to some parameter estimation problems*, in Distributed Parameter Systems, Springer Lecture Notes in Control and Inform. Sci., 75, pp. 1–23. Springer-Verlag, Berlin, New York, 1985.
- [3] H. T. BANKS AND K. ITO, *A unified framework for approximation in inverse problems for distributed parameter systems*, Control Theory Adv. Tech., 46 (1988), pp. 73–90.
- [4] H. T. BANKS AND F. KOJIMA, *Boundary shape identification problems in two-dimensional domains related to thermal testing of materials*, Tech. Report 88–23, ICASE, March 1988, Quart. Appl. Math., submitted.
- [5] H. T. BANKS AND K. KUNISCH, *Estimation techniques for distributed parameter systems*, CCS Lecture Notes, Birkhäuser, Boston, Providence, RI, 1989.
- [6] H. T. BANKS AND P. K. LAMM, *Estimation of variable coefficients in parabolic distributed systems*, IEEE Trans. Automat. Control, AC30 (1985), pp. 386–398.
- [7] H. T. BANKS, S. REICH, AND I. G. ROSEN, *An approximation theory for the identification of nonlinear distributed parameter systems*, SIAM J. Control Optim., submitted.
- [8] ———, *Galerkin approximation for inverse problems for nonautonomous nonlinear distributed systems*, Applied Math. Optim., submitted.
- [9] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1979.
- [10] M. DELFOUR, G. PAYRE, AND J.-P. ZOLÉSIO, *Optimal design of a minimum weight thermal diffuser with constraint on the output thermal power flux*, Applied Math. Optim., 9 (1983), pp. 225–262.
- [11] M. C. DELFOUR AND J.-P. ZOLÉSIO, *Shape sensitivity analysis via min max differentiability*, SIAM J. Control Optim., (1988), pp. 834–862.
- [12] R. F. DRENICK AND F. KOZIN, EDS., *System Modelling and Opt.*, in Proc. 10th IFIP Conf., Lecture Notes in Control and Inform. Sci., Vol 38, Springer-Verlag, Berlin, New York, August 1981.
- [13] S. GUTMAN AND L. W. WHITE, *On the estimation of l^∞ diffusion coefficients in parabolic equations*, to appear.
- [14] E. J. HAUG AND J. S. ARORA, *Applied Optimal Design*, Wiley-Interscience, New York, 1979.
- [15] E. J. HAUG AND J. CÉA, EDS., *Optimization of Distributed Parameter Structures*, Vols. I and II, Sijthoff and Noordhoff, Alphen aan den Rijn, the Netherlands, 1981.
- [16] E. J. HAUG, K. K. CHOI, AND V. KOMKOV, *Design Sensitivity Analysis of Structural Systems*, Academic Press, New York, 1986.
- [17] P. K. LAMM, *Estimation of discontinuous coefficients in parabolic systems: Applications to reservoir simulation*, SIAM J. Control Optim., 25 (1987), pp. 18–37.
- [18] ———, *Isoparametric finite element methods to estimate discontinuous coefficients in two-dimensional elliptic equations*, in Proc. 26th IEEE Conf. Decision and Control, pp. 1405–1410, Los Angeles, CA, December 1987.
- [19] P. K. LAMM, C. K. LO, AND I. G. ROSEN, *Identification of degenerate distributed parameter systems*, in IFAC Symposium on Control of Distributed Parameter Systems, Perpignan, France, June 1989.

- [20] P. K. LAMM AND K. A. MURPHY, *Estimation of discontinuous coefficients and boundary parameters for hyperbolic systems*, Quart. Appl. Math., 46 (1988), pp. 1–22.
- [21] P. K. LAMM AND I. G. ROSEN, *An approximation theory for the estimation of parameters in degenerate Cauchy problems*, J. Mathematical Analysis and Applications, 162 (1991), pp. 13–48.
- [22] P. LANCASTER AND M. TISMENETSKY, *The Theory of Matrices, Second Edition, with Applications*, Academic Press, New York, 1985.
- [23] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, New York, 1983.
- [24] O. PIRONNEAU, *Optimal Shape Design for Elliptic Systems*, Springer-Verlag, Berlin, New York, 1983.
- [25] R. E. SHOWALTER, *Hilbert Space Methods for Partial Differential Equations*, Pitman, Boston, 1979.
- [26] H. TANABE, *Equations of Evolution*, Pitman, Boston, 1979.
- [27] P. THOFT-CHRISTENSEN, ed., *System Modelling and Opt.*, in Proc. 11th IFIP Conf., Lecture Notes in Control and Inform. Sci., Vol 59, Springer-Verlag, Berlin, New York, July 1983.

SOME PROPERTIES OF DISTRIBUTED CONTROL SYSTEMS WITH FINITE-DIMENSIONAL INPUT SPACE*

LUCIANO PANDOLFI†

Abstract. This paper considers the “algebraic” Riccati operator equation, which corresponds to the quadratic regulator problem for a distributed control system (it is not assumed that the operators in the cost index are positive). The assumptions used in this paper are practically important: the input space is finite-dimensional and the pair (A, C) is modally observable (with respect to a suitable region Ω). If both these conditions hold, it is proved that the spectrum of the operator A in $\dot{x} = Ax + Bu$ must be quite a special set, whose structure depends on the properties of the solutions to the Riccati equation. The properties of $\sigma(A)$ derived in this paper must be compared with those derived in [C. A. Jacobson and C. N. Nett, *IEEE Transactions on Automatic Control*, AC-33 (1988), pp. 541–549] under the stabilization assumption.

Key words. linear systems, distributed systems, Riccati equation

AMS(MOS) subject classifications. 93C25, 93D15

1. Introduction and preliminary references. In this paper we consider a linear infinite-dimensional system over a complex Hilbert space X . The system is identified by a linear operator A from X to X , which is closed and densely defined, and by a linear and bounded operator $B, B \in L(U, X)$. Moreover, an operator $C, C \in L(X, Y)$ is given. The spaces X, Y, U are complex Hilbert spaces and *the crucial assumption is that U is finite-dimensional*.

Usually, in the applications to systems theory, the triple of operators (B, A, C) identifies the control system

$$\begin{aligned} (1) \quad & \dot{x} = Ax + Bu, \\ (2) \quad & y = Cx. \end{aligned}$$

In most of the cases, the operator A is the infinitesimal generator of a C_0 -semigroup on X , but, in recent times, optimal control problems for not well posed systems have been considered [8], [9], [19], [30], and, for this reason, we explicitly state the assumption that A is a generator when this assumption is really needed. We see that many properties can also be proved when A is not a generator, but it is more important that we are able to explain one of the major difficulties with the regulator problem for inverse systems: *the relevant solution to the Riccati equation is not bounded*. In §4 we see, in fact, that the Riccati equation cannot have bounded solutions when the spectrum of the operator A contains an unbounded sequence in $\{\operatorname{Re} z > 0\}$ when an “observability property” holds. The observability property is a main concern in this paper and is discussed in detail in the next section.

We explicitly note that the assumption that the operators B, C are bounded eliminates the important case of boundary control and observation. Some remarks on this case can be found in the final section.

*Received by the editors October 18, 1989; accepted for publication April 5, 1991.

†Politecnico di Torino, Dipartimento di Matematica Corso Duca degli Abruzzi, 24, 10129 Torino, Italy. The author was supported by the Italian Ministero della Ricerca Scientifica e Tecnologica, within the programs of Gruppo Nazionale per L'Analisi Funzionale ed Applicazioni–Consiglio Nazionale delle Ricerche.

As was stated, we assume that $\dim U < \infty$, a practically unavoidable assumption. In many applications, also, the output space Y is finite-dimensional. The case $\dim Y = +\infty$, however, is important, too: for example, the case where $Y = X$ is studied to construct (static) stabilizing feedbacks. For this reason, we do not make any assumption on the space Y , and we try to understand the limitations imposed by the finite-dimensionality of the input space. Among the oldest results along these lines, we quote [5], where some special cases of the properties in §3 have been derived, and [34]. In the latter, it is proved that exact controllability is impossible (with locally L^p -controls, $p > 1$) if U is finite-dimensional. This result must be contrasted with [38]. It is proved in [38] that exact controllability is equivalent to the property of full stabilization when the operator A generates a C_0 -group of operators. Indeed, more recent results on this subject were mostly concerned with the stabilization property ([4], [15], and references therein). An old sufficient condition for stabilization is the existence of a direct sum decomposition $X = X_- \oplus X_+$ with the following properties: (i) $\dim X_+ < \infty$ and A , the generator of a C_0 -semigroup, is reduced by this decomposition of X , $A = \text{diag}(A_-, A_+)$. Accordingly, we write $B = \text{col}(B_-, B_+)$; and (ii) the operator A_- generates an exponentially stable semigroup while (A_+, B_+) is a (finite-dimensional) stabilizable pair; i.e., there exists a matrix F such that $e^{(A_+ + B_+ F)t}$ is exponentially stable (see [33]).

An important result that can be found in the above-mentioned papers is that, for *detectable* systems with finite-dimensional space U , the above property is also a necessary condition. We recall that the pair (A, C) is detectable when there exists a bounded operator L , $L: Y \rightarrow X$, such that $(A + LC)$ generates an exponentially stable semigroup.

Remark. In [15] the assumption $\dim(Y) < \infty$ was also used.

Exponential stabilizability implies that the infinite horizon regulator problem is solvable. Conversely, it is proved in [37] that solvability of this last problem, plus detectability, implies that the pair (B, A) is stabilizable.

It is known that both the infinite horizon regulator problem and the stabilization problem are equivalent to the existence of positive solutions to a suitable Riccati equation [37]. For this reason we make the following assumption.

Assumption 1. The Riccati equation

$$(RE) \quad \langle Ax, Py \rangle + \langle x, PAy \rangle - \langle x, Ky \rangle = -\langle Cx, Cy \rangle \quad \forall x, y \in \text{Dom}(A)$$

admits a solution $P = P^*$. In (RE), K is a compact operator from X to X , $K = PHP$, and $H = H^* \in L(X)$.

We note that the assumption that P is positive will be declared when it is really needed.

In most of the application, H has the form $H = BB^*$, a positive operator. We do not assume from the outset that H is positive in this paper, since in some applications (conflicting control, or the important case of H^∞ control) we have that $K = P(BB^* - DD^*)P$, where D is a second finite-dimensional operator [29]. (See [21] for the case of distributed systems.) Consequently, in the following, the operator B will be subsumed in the operator H , and the finite-dimensionality of the input space U is replaced by the following assumption.

Assumption 2. The operator H that appears in (RE) is compact.

As we noted, the properties of “observability” or “detectability” will play a central role in our considerations. In fact, we make use of a frequency domain property, called Ω -modal observability, which is introduced and discussed in the next section.

The main contribution of the present paper is in the analysis of the interplay among modal observability, Riccati equation, and compactness of the operator H . In particular, we intend to stress the role of modal observability, which is discussed in the next section, with several examples to show the relations with the better-known observability and detectability properties. The main results are presented in §§3 and 4. For more clarity, we examine the case where $K = 0$ in §3. In this case, (RE) is reduced to the *Lyapunov equation*

$$(LE) \quad \langle Ax, Py \rangle + \langle x, PAy \rangle = -\langle Cx, Cy \rangle \quad \forall x, y \in \text{Dom} A.$$

Some of the results are already known in this case and are stated for completeness. The general case of (RE) is examined in §4. The final results of §4, Theorems 15–17, show, in particular, that in many cases that are important for the applications, modal observability and detectability are equivalent properties.

Some of the results in §4 are known for detectable systems.

2. The modal observability property. We discuss in this section the property of Ω -modal observability. We note that this property proves to have a special role in the stabilization problem and in the analysis of the relationships between internal and external description of distributed systems [2], [22]–[28].

DEFINITION. Let $\Omega \subseteq \mathbb{C}$ be a region. The pair (A, C) is *Ω -modally observable* when there exist holomorphic bounded operator-valued functions $X(z)$, $Y(z)$ (defined in Ω) such that

$$(3) \quad x = X(z)(zI - A)x + Y(z)Cx \quad \forall x \in \text{Dom}(A).$$

If z_0 is a complex number, we say that the pair (A, C) is *z_0 -modally observable* when it is Ω -modally observable with Ω some neighborhood of z_0 . We say that (A, C) is *uniformly Ω -modally observable* when $X(z)$, $Y(z)$ are bounded on Ω .

The most important case is, of course, the case where Ω is a half-plane, $\Omega = \{z, \Re z > -\alpha\}$. If A generates an exponentially stable C_0 -semigroup $E(t)$ (i.e., if $|E(t)| < Me^{-\sigma t}$, $\sigma > 0$), then $Y(z) = 0$ and $X(z) = (zI - A)^{-1}$ satisfy (3) for $\Re z > -\sigma + \epsilon$, for any $\epsilon > 0$. If the pair (A, C) is detectable, i.e., if there exists L such that $(A + LC)$ generates an exponentially stable semigroup, then $X(z) = (zI - A - LC)^{-1}$, $Y(z) = (zI - A - LC)^{-1}L$ satisfy (3) and are bounded over $\Re z > -\delta$ for some positive δ .

To clearly illustrate the interest of the above definition, let us assume for a moment that X is finite-dimensional. Moreover, let us introduce the notation $\Pi_\delta = \{z, \Re z > -\delta\}$ so that $\Pi_0 = \{z, \Re z > 0\}$. When Ω is all the complex plane, it is known that Ω -modal observability is equivalent to observability, and, if $\Omega = \Pi_0$, it is equivalent to detectability [11]. Moreover, in the first case, $X(z)$, $Y(z)$ can be chosen to be polynomial matrices, while, in the second case, they can be chosen to be proper stable rational matrices; in general, in the finite-dimensional case, $X(z)$, $Y(z)$ can be chosen to be rational matrices regular over Ω when this set is not the extended complex plane. For this reason, equality (3) is called a *Bézout equation*, or the *right coprimeness* property of C and $zI - A$, over the ring of rational functions that are regular on Ω . The property of modal observability is also known as *spectral observability* because of its geometrical interpretation (see [2]). Hence, property (3) and its “dual” property of *modal controllability* provide the link between the state space and the algebraic approach to linear time-invariant systems. They have a crucial role for many problems

connected with the frequency domain description of this class of systems, both in continuous and discrete time.

The books [3], [13], [14], [31] are largely devoted to studying the consequences of property (3) (and of “modal controllability”).

Let us go back to consider distributed parameter systems now. Also, in this case, property (3) proves to be useful in the analysis of the internal and external description of important classes of systems and for the stabilization problem, as we already have stated. Now, however, the relations between the properties of modal observability, observability, and detectability are much more difficult. The following examples illustrate this point.

The first is an easy example, with an operator that does not generate a semigroup.

Example 1. The space X is $L^2(0, 1)$, $C = 0$, and A is

$$\text{Dom} A = \{x \in W^{1,2}, x(0) = x(1) = 0, \} \quad Ax = -\dot{x}.$$

The operator A is closed and densely defined. It is easily seen that $\sigma(A) = \sigma_r(A) = \mathbf{C}$. Despite this, the holomorphic operator-valued function $z \rightarrow X(z)$

$$X(z)y = \int_0^t e^{-z(t-s)}y(s) ds$$

is a left inverse of $(zI - A)$. Even more, $\|X(z)\|$ is bounded on every half-plane $\Re z > \sigma$ for each number σ .

The next example is more involved, but it shows that even when A generates a semigroup, Π_0 -modal observability is much weaker than observability and detectability; in particular, it may hold with $C = 0$.

Example 2. It is well known that operators A exist, with the following properties:

(i) A is the generator of an exponentially stable semigroup; and (ii) the resolvent set of A contains Π_δ , $\delta > 0$ (see [36]). If the operator A has these properties, then $X(z) = (zI - A)^{-1}$ is holomorphic on Π_0 , so that Π_0 -modal observability holds also with $C = 0$.

Even more concrete examples can be given.

Example 3. In this example, $\Omega = \mathbf{C}$ and the system is modally observable, despite the fact that it is not observable (the analogous case is impossible for finite-dimensional systems). The example is suggested by the considerations in [22], [23], where further interesting observations can be found. It is described by

$$\frac{d}{dt} \begin{bmatrix} x(t) \\ \xi(t) \end{bmatrix} = \begin{bmatrix} \xi(t) \\ x(t-1) - \xi(t-1) \end{bmatrix}, \quad y(t) = \xi(t).$$

It is known that this system can be represented in the form (1), (2), and that modal observability is equivalent to

$$\ker \begin{pmatrix} z & -1 \\ -e^{-z} & z + e^{-z} \\ 0 & 1 \end{pmatrix} = \{0\}$$

(see the references quoted above). The condition for modal observability holds, but the system is not observable, since the nonzero initial condition $(x(t), \xi(t))$, $x(\vartheta) = \sin 2\pi\vartheta = \xi(\vartheta)$ for $-1 \leq \vartheta \leq 0$ gives zero output. See [23, §5] for more interesting examples.

Finally, the following example from [20] also deserves to be reported.

Example 4. The example is described by the following equation:

$$\frac{d}{dt} \begin{bmatrix} x(t) \\ \xi(t) \end{bmatrix} = \begin{bmatrix} -x(t) + u(t) \\ \frac{d}{dt}\xi(t-1) + x(t) - \xi(t) \end{bmatrix}, \quad y(t) = \xi(t).$$

This system can be written in the form (1), and it is easily seen that it is both modally observable and controllable. Moreover, its transfer function is holomorphic on Π_0 and bounded on the closure of (Π_0) , so that the system is input-output stable. Despite this, its free evolutions are not exponentially stable, a fact that cannot happen in finite dimensions.

Example 4 is examined again at the end of §4.

The previous examples show that modal observability is a very weak property. However, experience with finite-dimensional systems and the papers already quoted suggest that it is one of the crucial properties of linear systems. Hence it is important both to understand the results that can be derived from it and to know conditions under which it is equivalent to the property of detectability.

3. Lyapunov Equation. We assume in this section that a solution P to (LE) exists (we need $P \in L(X)$, but, for the moment, P need not be either positive or symmetric). For more clarity, we first consider the spectral properties of the operator A at $z_0 = 0$. We assume a simpler form than 0-modal observability for the first results: namely, we assume that there exist two operators X, Y such that

$$(4) \quad XAx + YCx = x \quad \forall x \in \text{Dom } A.$$

The equivalence of condition (4) and 0-modal observability was proved in [26, Thm. 2].

It is easily seen that (i) if $\{x_n\}$ is a bounded sequence in the domain of A such that $Ax_n \rightarrow 0$, then $Cx_n \rightarrow 0$, by putting $x = y = x_n$ in (LE); and (ii) if $\{x_n\}$ is a bounded sequence in the domain of A such that $Ax_n \rightarrow y_0$, then $\{Cx_n\}$ is a Cauchy sequence (put $x = y = (x_n - x_m)$ in (LE)). In particular, if the operator C^*C has a bounded inverse, $\{x_n\}$ is convergent, $x_n \rightarrow x_0$. Then $x_0 \in \text{Dom } A$ and $Ax_0 = y_0$, since A is a closed operator.

LEMMA 1. *Let us assume that condition (4) holds and that a solution P exists to (LE). Then any sequence $\{x_n\}$ such that $\{Ax_n\}$ is convergent is itself convergent.*

Proof. We first assume that $\{x_n\}$ is bounded. It is shown above that $\{Cx_n\}$ is convergent. Then the result follows from the equality (4) with $x = x_n$:

$$x_n = XAx_n + YCx_n.$$

Consequently, $\{x_n\}$ converges to some x_0 , and $Ax_0 = \lim Ax_n$, since the operator A is closed. Now we show that $\{x_n\}$ is bounded. Let us assume, by contradiction, that this is not the case. Then we put $\eta_n = x_n/\|x_n\|$, and we see that $A\eta_n \rightarrow 0$ (and $\{\eta_n\}$ is bounded) so that $C\eta_n \rightarrow 0$. Condition (4) implies that $\{\eta_n\}$ tends to zero. This is impossible, since $\|\eta_n\| = 1$ for each index n . \square

COROLLARY 2. *Under the same assumptions as those in Lemma 1, $\text{Im } A$ is a closed subspace of X .*

Proof. In fact, it is shown that if $Ax_n \rightarrow \xi$, then $x_n \rightarrow x_0$. As A is a closed operator, then $x_0 \in \text{Dom } A$ and $Ax_0 = \xi$. \square

THEOREM 3. *Under the same assumptions as those in Lemma 1, $0 \notin \sigma_p(A) \cup \sigma_c(A)$.*

Proof. The point $z_0 = 0$ is not in the continuous spectrum, since the closed operator A has closed image; it is not in the point spectrum, since if $Ax_0 = 0$ and $x_0 \neq 0$, then, from (LE) with $x = y = x_0$, we see that $Cx_0 = 0$ and, from (4), x_0 should be the null vector. \square

The previous result asserts that we should not assume that 0 is not a point in $\sigma_r(A)$; in fact, it may well be that $0 \in \sigma_r(A)$, as the following example shows.

Example 5. Let $X = l^2$ and A be the right shift, $A(x_i) = (y_i)$, $y_1 = 0$, $y_i = x_{i-1}$ $i \geq 2$. Let $P = (A + A^*)$ so that $(-A^*)P + P(-A) = -(2I + (A^*)^2 + A^2)$ is negative. The operator C is the square root of $(2I + (A^*)^2 + A^2)$. Then (LE) is satisfied by P , and condition (4) holds with $X = A^*$, $Y = 0$, since $A^*A = I$. Despite this, $0 \in \sigma_r(A)$.

The analysis of the spectral properties of the operator A at $z_0 = 0$ is finished by the following result.

THEOREM 4. *Let us assume that 0-modal observability holds and that there exists a solution P to (LE). Then $0 \notin \text{cl} \{\sigma_p(A) \cup \sigma_c(A)\}$.*

Proof. Otherwise, we can find sequences $\{x_n\}$, $\{z_n\}$, $x_n \in X$, $\|x_n\| = 1$, $z_n \in \mathbb{C}$, $z_n \rightarrow 0$, such that $(z_n I - A)x_n \rightarrow 0$. In fact, if $z_n \in \sigma_p(A)$, then x_n is one of the eigenvectors of z_n ; if $z_n \in \sigma_c(A)$, then $(z_n I - A)$ is not invertible and we can find x_n such that $\|(z_n I - A)x_n\| < (1/n)$. From (LE),

$$\begin{aligned} & \langle Cx_n, Cx_n \rangle + z_n \langle x_n, Px_n \rangle + \bar{z}_n \langle Px_n, x_n \rangle \\ &= \langle (z_n I - A)x_n, Px_n \rangle + \langle Px_n, (z_n I - A)x_n \rangle \end{aligned}$$

so that $Cx_n \rightarrow 0$. We recall that 0-modal observability means that equality (3) holds for z in a neighborhood of $z_0 = 0$. Hence, by putting $z = z_n$, $x = x_n$ in (3), we see that $x_n \rightarrow 0$. This is impossible, since the norm of x_n is equal to 1 for each index n . \square

Remark. We note that an analogous proof allows us to conclude that the image of A is closed; this is a property that the point 0 must have even if it belongs to $\sigma_r(A)$. An analogous statement, which is not explicitly quoted, also holds for the results that we will now prove.

We now examine the spectrum of A in a strip around the imaginary axis.

THEOREM 5. *Let us assume that (LE) has a solution $P = P^*$ and that uniform Ω -modal observability holds over $\Omega = \{z, |\Re z| < \alpha\}$. Then there exists $\sigma > 0$ such that $\{\sigma_p(A) \cup \sigma_c(A)\} \cap \{z, |\Re z| < \sigma\} = \emptyset$.*

Proof. Otherwise, there exist a sequence z_n of complex numbers and a sequence x_n of unitary vectors in X such that $(z_n I - A)x_n \rightarrow 0$ and $\Re z_n \rightarrow 0$. Hence, from (LE),

$$(5) \quad \|Cx_n\|^2 + 2(\Re z_n) \langle x_n, Px_n \rangle = 2\Re \langle (z_n I - A)x_n, Px_n \rangle$$

and $Cx_n \rightarrow 0$. Uniform modal observability implies that $x_n \rightarrow 0$, while we assumed that the norm of x_n is one. \square

Finally, we consider the case of a half-plane. If δ is any real number, we recall that $\Pi_\delta = \{z, \Re z > -\delta\}$.

THEOREM 6. *We assume that*

- (i) *The conditions of Theorem 5 hold;*
- (ii) *(LE) has a solution $P = P^* > 0$;*
- (iii) *There exists modal observability over $\Pi_0 = \{z, \Re z > 0\}$.*

Then there exists a number $\sigma > 0$ such that $\{\sigma_p(A) \cup \sigma_c(A)\} \cap \Pi_\sigma = \emptyset$. Moreover, $zI - A$ has a continuous left inverse for each z with $\Re z > 0$.

Proof. We show that if $z \in \sigma_p(A) \cup \sigma_c(A)$, then $\Re z$ cannot be positive. This will follow from P being positive. This is sufficient, since we already know that no point of $\sigma_p(A) \cup \sigma_c(A)$ belongs to $-\sigma < \Re z < 0$ from Theorem 5. If $z \in \sigma_p(A) \cup \sigma_c(A)$, then there exists a sequence $\{x_n\}$ of unitary vectors such that $(zI - A)x_n \rightarrow 0$ ($\{x_n\}$ may be a stationary sequence if $z \in \sigma_p(A)$). In this case, $x_n \equiv x_0 \in \ker(zI - A)$. Hence the right-hand side of (5), with $z_n = z$, tends to zero. As $P = P^* \geq 0$, we see that $\langle x_n, Px_n \rangle \geq 0$, so that $x_n \rightarrow 0$ and $Cx_n \rightarrow 0$. This contradicts modal observability.

In particular, we have seen that $\ker(zI - A) = \emptyset$ for each z , $\Re z > 0$, so that the left inverse of $zI - A$ exists. If, for some z_0 , the inverse is not continuous, then there exists a sequence $\{x_n\}$ of vectors of unitary norms such that $(z_0I - A)x_n \rightarrow 0$. The above arguments show that this is impossible. \square

In the proof of Theorem 6 we have explicitly proved the following result, which will be used later.

COROLLARY 7. *Under the assumptions (2), (3) of Theorem 6, $\{\sigma_p(A) \cup \sigma_c(A)\} \cap \Pi_0 = \emptyset$, and $(zI - A)$ has a bounded left inverse on Π_0 .*

The previous theorem completes the analysis of the spectral properties of A if the operator A is selfadjoint, since, in this case, $\sigma_r(A)$ is empty. Otherwise, we have the following results.

THEOREM 8. *Let us assume that A generates a C_0 -semigroup $E(t)$. If $C = I$ and if (LE) has a solution $P = P^* \geq 0$, then there exists a number $\sigma > 0$ such that $\sigma(A) \subseteq \{z, \Re z < -\sigma\}$.*

Proof. From [6] it is known that $E(t)$ is exponentially stable. \square

In recent times, “stable” Bézout equations proved to be important tools. Knowing this fact, we state the following result, which extends Theorem 8. We recall that $H^\infty(\Pi_\delta)$ denotes the Banach space of those functions, which are holomorphic and bounded over Π_δ ; H^∞ denotes $H^\infty(\Pi_0)$. Hence uniform modal controllability on Π_0 means that the functions $X(\cdot)$, $Y(\cdot)$ belong to H^∞ . The Hardy space $H^2 = H^2(\Pi_0)$ has a more complex definition. For the following result, it is sufficient to know that it is the space of the Laplace transforms of functions that are square integrable on $(0, +\infty)$; the Laplace transform is an algebraic and topological isomorphism between $L^2(0, +\infty)$ and H^2 .

THEOREM 9. *Let us assume that*

- (i) A generates a C_0 -semigroup $E(t)$;
- (ii) uniform modal observability holds on Π_0 ;
- (iii) (LE) admits a solution $P = P^* \geq 0$.

Under these assumptions, the semigroup $E(t)$ is exponentially stable, so that there exists $\sigma > 0$ such that $\sigma(A) \cap \{z, \Re z > -\sigma\} = \emptyset$.

Proof. It is sufficient to show that under the stated assumptions the semigroup $E(t)$ is exponentially stable.

Let $x \in$ belong to the domain of A . Then

$$\begin{aligned} -\|CE(t)x\|^2 &= \langle AE(t)x, PE(t)x \rangle + \langle PE(t)x, AE(t)x \rangle \\ &= \frac{d}{dt} \langle E(t)x, PE(t)x \rangle. \end{aligned}$$

Integration then shows that

$$\int_0^T \|CE(t)x\|^2 dt \leq \|Px\|^2 - \langle E(T)x, PE(T)x \rangle \leq \|Px\|^2.$$

Of course, this inequality also holds for any $x \in X$, since $\text{Dom } A$ is dense in X . Moreover, we can put $T = \infty$ and still have the same estimate. Hence the function $CE(t)x$ is a square integrable function. In particular, its Laplace transform $\Phi(z)x$ belongs to H^2 , its H^2 -norm being less than $\kappa\|x\|$ for a positive number κ .

From the properties of C_0 -semigroups, it is known that there exists a halfplane Π_r over which the operator $(zI - A)^{-1}$ exists as a bounded operator. It is not restrictive to assume that $r < 0$. On Π_r the Laplace transform $\Phi(z)x$ of $CE(t)x$ is given by $\Phi(z)x = C(zI - A)^{-1}x$; on Π_r , from condition (2),

$$\begin{aligned}(zI - A)^{-1}x &= X(z)x + Y(z)C(zI - A)^{-1}x \\ &= X(z)x + Y(z)\Phi(z)x \quad x \in \text{Dom } A.\end{aligned}$$

Hence $(zI - A)^{-1}$ has a holomorphic extension to Π_0 , which must coincide with the left inverse, whose existence is asserted in Corollary 7. Let us denote by $\Delta(z)$ this extension. The holomorphic function $\Delta(z)$ is the sum of an H^∞ function (namely, $X(z)$) and of an H^2 function (namely, $Y(z)\Phi(z)$). By using an idea in [35], we see that this implies that the semigroup $E(t)$ is exponentially stable.

Let $F(z)$ be the function $F(z)x = \int_0^1 e^{-zt}E(t)x \, dt$, so that $F(\cdot) \in H^2 \cap H^\infty$ and

$$\sup_{\{\text{Re } z > 0\}} \|F(z)x\| \leq \int_0^1 \|E(t)x\| \, dt \leq \alpha\|x\|.$$

Hence $\Delta(z)F(z)x = X(z)F(z)x + Y(z)\Phi(z)F(z)x$ is in H^2 , and its H^2 -norm is bounded uniformly with respect to x for $\|x\| \leq 1$. In fact,

- (i) $\|X(z)F(z)\|_{H^2} \leq \|X(z)\|_{H^\infty} \times \|F(z)\|_{H^2}$;
- (ii) $\|Y(z)C(zI - A)^{-1}F(z)\|_{H^2} \leq \|Y(z)\|_{H^\infty} \times \|\Phi(z)\|_{H^2} \times \|F(z)\|_{H^\infty}$.

Consequently, $z \rightarrow \Delta(z)F(z)x$ belongs to H^2 ; i.e., its inverse Laplace transform is a square integrable function. This inverse Laplace transform is

$$\begin{aligned}tE(t)x &\text{ for } 0 \leq t < 1, \\ E(t)x, &\quad t \geq 1\end{aligned}$$

because $\Delta(z)F(z)x = (zI - A)^{-1}F(z)x$ on Π_r . Hence the semigroup generated by A is L^2 -, i.e., exponentially stable [6]. \square

Remark. The assumption that A is a generator is essential for the previous result. In fact, Example 1 shows an operator A (which is not a generator) with the following property: $(zI - A)$ has a left inverse that belongs to $H^\infty(\Pi_0)$, so that uniform modal observability holds with $C = 0$; if $C = 0$, then (LE) has the solution $P = 0$.

4. The Riccati equation. We now consider the Riccati equation (RE). In the first results (Lemmas 10–12), we simply assume that $K = K^*$, but we do not assume that K has the special form $K = PHP$. We note that, for bounded A and any $P = P^*$, we can always find K such that P is a solution of the corresponding (RE): it will be $K = A^*P + PA + C^*C$ and K is compact if P, C are compact operators. So the properties of $\sigma(A)$ that can be derived in this case are very weak, and essentially depend upon modal observability of the system.

We first prove a lemma.

LEMMA 10. *Let us assume that z_0 is a complex number with the following properties:*

- (i) $\ker(z_0I - A) = \{0\}$;
- (ii) $\inf\{\|(z_0I - A)x\|, x \in \text{Dom } A, \|x\| = 1\} = 0$.

Let V be a finite-dimensional subspace of the domain of A . Then

$$\inf\{\|(z_0I - A)x\|, x \in \text{Dom } A, \|x\| = 1, x \in V^\perp\} = 0.$$

Proof. Let us assume, by contradiction, that the infimum over V^\perp is positive. In this case, there exists a bounded left inverse D of $(z_0I - A)|_{V^\perp}$, $\text{Dom } D = \text{cl Im}(z_0I - A)|_{V^\perp}$.

By assumption, we can find a normalized sequence $\{x_n\}$ in $\text{Dom } A$ such that $(z_0I - A)x_n \rightarrow 0$. We represent $x_n = v_n + w_n$, $v_n \in V$, $w_n \in V^\perp$. As $\|v_n\| \leq 1$ and V is locally compact, we can assume that $v_n \rightarrow v_0$, $(z_0I - A)v_n \rightarrow (z_0I - A)v_0$, so that $\{(z_0I - A)w_n\}$ is also convergent. Since $(z_0I - A)w_n \in \text{Im}(z_0I - A)|_{V^\perp}$, then w_n is convergent, say $w_n \rightarrow w_0$. The operator A being closed, $(z_0I - A)w_n \rightarrow (z_0I - A)w_0$ and $(z_0I - A)(v_0 + w_0) = 0$. From the first assumption, $v_0 + w_0 = 0$, in contrast with the assumption that $1 = \lim \|x_n\| = \|v_0 + w_0\|$. \square

Now we can consider the purely imaginary points in $\sigma(A)$. We recall the standing assumption that $K = K^*$ is a compact operator, but still we do not require that $K = PHP$.

LEMMA 11. *Let $i\omega$ be a fixed number. We assume that $i\omega$ -observability holds. Then it is not possible to find a sequence $\{x_n\}$ in the domain of A such that $\|x_n\| = 1$, $x_n \rightarrow 0$ and $(i\omega - A)x_n \rightarrow 0$.*

Proof. Let $\xi_n = (i\omega - A)x_n$. Then, from (RE) with $x = y = x_n$, we have that

$$-2\Re\langle \xi_n, Px_n \rangle - \langle x_n, Kx_n \rangle = -\|Cx_n\|^2$$

and the left-hand side converges to zero: $Kx_n \rightarrow 0$, since $x_n \rightarrow 0$ and K is a compact operator; $\langle \xi_n, Px_n \rangle \rightarrow 0$, since $\{Px_n\}$ is bounded and $\xi_n \rightarrow 0$. Hence $Cx_n \rightarrow 0$, so that, from $i\omega$ -modal observability,

$$x_n = X(i\omega)(i\omega - A)x_n + Y(i\omega)Cx_n,$$

so that $x_n \rightarrow 0$. This is not possible, since we assumed that the norm of x_n is equal to 1 for every n . \square

The previous lemma shows that the following possibilities are inconsistent with $i\omega$ -modal observability and compactness of K :

- (i) $i\omega \in \sigma_p(A)$ and $\dim \ker(i\omega I - A) = +\infty$. In fact, in this case, we can find an infinite orthonormal sequence $\{x_n\}$ in $\ker(i\omega I - A)$. Any infinite orthonormal sequence $\{x_n\}$ converges weakly to zero, from the Riesz–Fischer theorem.
- (ii) $i\omega \in \sigma_c(A)$. In fact, let $x_1, \|x_1\| = 1$ be such that $\|(i\omega I - A)x_1\| < 1$. Once that x_2, \dots, x_n have been chosen, we can find

$$x_{n+1} \in \text{span}\{x_1, \dots, x_n\}^\perp, \|x_{n+1}\| = 1,$$

so that $\|(i\omega I - A)x_{n+1}\| < 1/(n+1)$ (from Lemma 1). Hence $(i\omega I - A)x_n \rightarrow 0$ and $x_n \rightarrow 0$ (since $\{x_n\}$ is a orthonormal sequence).

- (iii) $i\omega \in \sigma_r(A)$, and $(i\omega I - A)$ does not have a bounded left inverse (by an analogous argument relying on Lemma 1).

Now let $\delta > 0$ and, as above, $\Pi_\delta = \{z, \Re z > -\delta\}$. An analogous proof results in the following lemma.

LEMMA 12. *The following properties are inconsistent:*

- (i) *There exists $P = P^*$, which solves (RE) (and $K = K^*$ is any compact operator);*
- (ii) *There exists $\delta \geq 0$ such that Π_δ -modal observability holds;*
- (iii) *There exists a sequence $\{z_n\}$, $\Re z_n \rightarrow \alpha \geq -\delta$ of complex numbers and a sequence $\{x_n\}$ of normalized vectors in $\text{Dom } A$ such that $x_n \rightarrow 0$, $(z_n I - A)x_n \rightarrow 0$*
- (iv) *The sequences $\{X(z_n)\}$, $\{Y(z_n)\}$ are bounded;*
- (v) *$\limsup \Re z_n \langle x_n, Px_n \rangle \rightarrow l$, $0 \leq l \leq +\infty$.*

Proof. We pass to a subsequence such that $\Re z_n \langle x_n, Px_n \rangle \rightarrow l$, $0 \leq l \leq +\infty$. We note that

$$\begin{aligned} & \langle (A - z_n I)x_n, Px_n \rangle + \langle Px_n, (A - z_n I)x_n \rangle - \langle x_n, Kx_n \rangle \\ &= -\|Cx_n\|^2 - 2\Re z_n \langle x_n, Px_n \rangle. \end{aligned}$$

As K is a compact operator, the left-hand side tends to zero. Hence $\Re z_n \langle x_n, Px_n \rangle \rightarrow 0$ (since $l \geq 0$) and $Cx_n \rightarrow 0$. However, $x_n = X(z_n)(z_n I - A)x_n + Y(z_n)Cx_n$, so that $x_n \rightarrow 0$ since the sequences $\{X(z_n)\}$, $\{Y(z_n)\}$ are bounded. This is a contradiction, since we assumed that $\|x_n\| = 1$. \square

In particular, this lemma is useful when $P \geq 0$ and $\Re z_n \rightarrow \alpha \geq 0$.

The possibilities that are eliminated by the previous lemma are (we put $\delta = 0$, this being the most important case)

- (i) That $z_0 \in \sigma_r(A)$, $\Re z_0 \geq 0$ and that $(z_0 I - A)$ does not have a bounded left inverse;
- (ii) That there exists $z_0 \in \sigma_c(A)$, $\Re z_0 \geq 0$.

In both these cases, in fact, we can put $z_n = z_0$ for each index n .

The sequence $\{x_n\}$ is constructed as in Lemma 1;

- (iii) $\sigma_p(A) \cap \{z, \Re z \geq 0\}$ contains an eigenvalue of infinite multiplicity.

Remark. An observation that is important to understanding the next example is the following one: we note that if $z_n \rightarrow z_0 \in \Pi_\sigma$, then $\{X(z_n)\}$ and $\{Y(z_n)\}$ are indeed bounded sequences. An analogous proof shows that if Ω is any region in $\Re z \geq 0$, then properties (i)–(iii) of Lemma 12 and (a) $z_n \rightarrow z_0 \in \Omega$; and (b) Ω -modal observability are inconsistent. The proof is analogous to that of Lemma 12.

Under the stated assumptions, we cannot prove more properties of $\sigma(A)$. In particular, it is not true that $\sigma_p(A) \cap \{z, \Re z \geq 0\}$ is finite, as the following example shows.

Example 6. Let $X = l^2$ and $U = \mathbf{C}$. The operator A is the left shift,

$$A(x_1, x_2, x_3, \dots) = (x_2, x_3, \dots),$$

while the operator C is defined by

$$C(x_1, x_2, x_3, \dots) = x_1.$$

We already noted that a solution P as in Lemma 12 exists, since A is a bounded operator.

As it is known, if $|z| < 1$, then $z \in \sigma_p(A)$. Despite this, the pair (A, C) is Ω -modally observable, $\Omega = \{z, |z| < 1\}$. In fact, with $u = (1, 0, 0, \dots)^*$, the operator-valued functions that satisfy (3) are

$$X(z) = -(I - zS)^{-1}S, \quad Y(z) = (I - zS)^{-1}u.$$

Here S is the right shift $S(x_1, x_2, x_3, \dots) = (0, x_1, x_2, \dots)$, so that $(I - zS)^{-1}$ exists for $|z| < 1$ as a bounded operator.

The operator A of this example is bounded, so that it is even the generator of a holomorphic group of operators.

Now we can modify this example so that it fits exactly in the framework in Lemma 12. It is known that $\zeta = 1$ belongs to $\sigma_c(A)$ and that the operator $\tilde{A} = (I - A)^{-1}(I + A)$ is densely defined and closed. Let $\tilde{C} = C$:

$$\tilde{X}(z) = X\left(\frac{z-1}{z+1}\right)\frac{(I-A)}{(z+1)}, \quad \tilde{Y}(z) = Y\left(\frac{z-1}{z+1}\right).$$

The functions $\tilde{X}(z)$, $\tilde{Y}(z)$ are holomorphic and bounded on Π_0 and $\sigma(\tilde{A}) \supseteq \Pi_0$. Moreover,

$$\tilde{X}(z)(zI - \tilde{A}) + \tilde{Y}(z)\tilde{C} = I \quad \text{for } \Re z > 0$$

since $|(z-1)/(z+1)| < 1$ if $\Re z > 0$.

Despite this, a symmetric solution P to (RE) exists, with some operator K (not of the form PHP): take $P = (I - A)^*Q(I - A)$ with $Q = Q^* \geq 0$ and $K = C^*C + (A + I)^*Q(I - A) + (I - A)^*Q(A + I)$.

Of course, we can obtain more stringent results if we introduce the assumption that $K = PHP$, $H = H^*$ a compact operator, which is a natural assumption from the point of view of systems theory. So this assumption will hold for the remainder of this paper.

LEMMA 13. *Under assumptions (i) and (ii) of Lemma 12 with $\delta = 0$, if $K = PHP$, H a compact operator, and $P \geq 0$, then it is not possible to find a sequence $\{z_n\} \in \sigma_p(A)$ such that*

- (i) *The sequence $\{Y(z_n)\}$ is bounded;*
- (ii) *$\limsup \Re z_n \geq 0$;*
- (iii) *The sequence $\{z_n\}$ is unbounded.*

Proof. Let us assume by contradiction that we can find sequences $\{z_n\}$, $\{x_n\}$ such that $Ax_n = z_nx_n$, $\|x_n\| = 1$, $\lim |z_n| = \infty$, and $\limsup \Re z_n \geq 0$. Then, for $y \in \text{Dom } A$,

$$z_n \langle x_n, Py \rangle = \langle Ax_n, Py \rangle = -\langle x_n, PAy \rangle + \langle x_n, PHPy \rangle - \langle Cx_n, Cy \rangle,$$

and the right-hand side is bounded for $n \rightarrow \infty$. If $|z_n| \rightarrow \infty$, then we must have $\langle x_n, Py \rangle = \langle Px_n, y \rangle \rightarrow 0$ for every y in the domain of A . Consequently, $Px_n \rightarrow 0$, since the domain of A is dense in X ; hence $PHPx_n \rightarrow 0$, since H is compact. Now, from (RE) with $x = y = x_n$,

$$2\Re z_n \langle x_n, Px_n \rangle + \|Cx_n\|^2 = \langle x_n, PHPx_n \rangle \rightarrow 0,$$

so that $Cx_n \rightarrow 0$, since the upper limits of both the terms on the right-hand sides are not negative. From modal observability, $x_n \rightarrow 0$. This is a contradiction, since $\|x_n\| = 1$. \square

If there exist infinitely many eigenvalues of A in $\Re z \geq 0$, then the relative eigenvectors must converge, as stated in the next result.

LEMMA 14. *Under the same assumptions as those in Lemma 13, let $\{z_n\}$ be a sequence in $\sigma_p(A) \cap \{z, \Re z \geq 0\}$. Let $\{v_n\}$ be a sequence of normalized eigenvectors, v_n relative to z_n . Then any weak limit point v_0 of $\{v_n\}$ is a strong limit point.*

Proof. Lemma 13 shows that it is not restrictive to assume that $z_n \rightarrow z_0$. As $\{v_n\}$ is bounded, we can assume that $v_n \rightharpoonup v_0$, so that $Av_n \rightharpoonup z_0 v_0$. An operator A is closed if and only if it is weakly closed, so that $v_0 \in \text{Dom } A$, $Av_0 = z_0 v_0$. Now, if $\{v_n\}$ does not tend to v_0 , there exists a positive number α such that $0 < \alpha < \|v_n - v_0\|$. Hence $x_n = (v_n - v_0)/\|v_n - v_0\| \rightarrow 0$, $(z_n I - A)x_n = (z_0 - z_n)v_0/\|v_n - v_0\| \rightarrow 0$, and this is in contrast with Lemma 12. \square

However, we cannot infer that $\sigma(A) \cap \Pi_\delta$ is finite for some δ . The example is still Example 1. Here we have an operator A whose resolvent has a holomorphic left inverse $X(z)$ for every z , and $\sup_{\Re z > \delta} \|X(z)\|$ is bounded for each δ . Hence uniform modal observability holds on $\Re z > \delta$, with $C = 0$. If $H = 0$, then $P = 0$ is a solution of (RE). Despite this, the resolvent set of A is empty.

The difficulties that arise when studying the regulator problem for inverse systems are partly explained by the previous lemmas. The infinite horizon regulator problem with positive operators for “inverse” systems is studied in [9]. As we can expect, its solution depends on (RE) (with $H = BB^*$) having a positive solution P . This is proved under the assumption that (with the symbols of the present paper) $(-A, B)$ is approximately controllable and $(-A, C)$ is detectable (the reason for these assumptions is the following: the solution is constructed as the inverse of the operator Q , which is the solution of the Riccati equation formally obtained for $Q = P^{-1}$. The result is that the solution Q exists under controllability, and that $\ker(Q) = \{0\}$ under detectability, so that $P = Q^{-1}$ is well defined). The result, however, is that the operator P is, in general, unbounded. The previous lemmas clearly indicate the reason: among the systems that fall in the approach in [9], there are systems with $C = I$ and the operator A , similar to the one that describes the inverse heat equation with zero Dirichlet conditions, whose eigenfunctions span the space X and correspond to a sequence of eigenvalues in Π_0 . The previous results imply that the operators P that solve (RE) and are positive cannot be bounded.

We now remain with the properties of $\sigma_r(A)$. As for the case of Lyapunov equation, we need some “stabilization” property, and we must require that A is a generator, after Example 1. We prove the following result, which also implies that A has at most finitely many unstable eigenvalues.

THEOREM 15. *We assume that*

- (i) *A is the infinitesimal generator of a C_0 -semigroup;*
- (ii) *There exists a solution P to (RE), $P = P^* > 0$, and $K = PHP$. The operator $H = H^*$ is compact;*
- (iii) *Π_0 -uniform modal observability holds.*

Then there exists a number $\gamma > 0$ such that $\sigma(A) \cap \Pi_\gamma$ is a finite set of eigenvalues, each one of finite multiplicity.

Proof. We note that (RE) can be written as

$$\langle (A + J)x, Py \rangle + \langle x, P(A + J)y \rangle = -\langle Cx, Cy \rangle - \langle HPx, HPy \rangle,$$

where $J = -(HP + H^*HP)/2$. We can view J as a bounded perturbation of A , so that $A + J$ generates a C_0 -semigroup that we denote $E'(t)$. Clearly, for $x \in \text{Dom } A$,

$$-\frac{d}{dt} \langle E'(t)x, PE'(t)x \rangle = \|CE'(t)x\|^2 + \|HPE'(t)x\|^2,$$

so that

$$\int_0^{+\infty} \|CE'(t)x\|^2 dt + \int_0^{+\infty} \|HPE'(t)x\|^2 dt \leq \langle x, Px \rangle \leq \|P\| \|x\|$$

for each $x \in X$, since $\text{Dom } A$ is dense in X . Consequently, the transformations $z \rightarrow C(zI - A - J)^{-1}x$, $z \rightarrow HP(zI - A - J)^{-1}x$ belong to the Hardy space H^2 . From the equality

$$\begin{aligned} (zI - A - J)^{-1}x &= X(z)x + X(z)J(zI - A - J)^{-1}x \\ &\quad + Y(z)C(zI - A - J)^{-1}x, \end{aligned}$$

we see that $(zI - A - J)^{-1}x$ has a holomorphic extension to Π_0 , which is the sum of a H^∞ and an H^2 function. As in Theorem 9, this implies that the semigroup $E'(t)$ is exponentially stable.

The operator A is a compact perturbation of $A + J$, so that it generates a quasicompact semigroup. Consequently, only one of the following possibilities may hold: (i) the set $\sigma(A) \cap \Pi_0$ is empty or (ii) the set $\sigma(A) \cap \Pi_0$ is finite and its elements are eigenvalues of finite multiplicity. (see [1, pp. 215–216]). In fact, let γ be minor than the exponential order of the semigroup $E'(t)$. Then also $e^{\gamma t}E'(t)$ is a quasicompact semigroup, whose generator is $(\gamma I + A + J)$. The above arguments can be repeated with $(\gamma I + A)$ in the place of A , and we get the result. \square

Let $\gamma > 0$ be as in Theorem 15 and assume that the set $\sigma(A) \cap \Pi_\gamma$ is not empty. Let N be the (finite-dimensional) generalized eigenspace of A with respect to the eigenvalues in Π_γ , and $P, I - P$ be the relative spectral projections. Then the following theorem holds.

THEOREM 16. *Let X_1 be the image of $I - P$ and A_1 be the operator $(I - P)A|_{X_1}$. Then A_1 generates an exponentially stable semigroup on X_1 .*

Proof. This follows from [1, p. 216]. \square

Finally, we consider the case where $H = BB^*$. Still with $N \neq \emptyset$, let A_N, B_N, C_N be the operators $PA|_N, PB, PC|_N$. The pair (B_N, A_N) identifies a linear finite-dimensional control system and the eigenvalues of A have nonnegative real parts.

THEOREM 17. *Under the assumptions of Theorem 15, the finite-dimensional system (B_N, A_N) is stabilizable, hence completely controllable, and the finite-dimensional system (A_N, C_N) is observable, hence detectable.*

Proof. In fact, the operator P is solution to the Lyapunov equation for the operator $A - \frac{1}{2}BB^*P$, and the assumptions of Theorem 9 are satisfied. Hence the pair (B, A) is stabilizable, and the conclusion follows, for example, from [11], since the spectrum of A_N lies in Π_0 . To see the second part, we represent the matrix A in the block form

$$A = \begin{pmatrix} A_1 & 0 \\ A_{2,1} & A_N \end{pmatrix}$$

and, correspondingly, $C = [C_1, C_N]$. If the finite-dimensional system identified by A_N and C_N is not observable, then $X_N(z)(zI - A_N) + Y_N(z)C = I$ cannot be solved with $X_N(z), Y_N(z)$ bounded in $\Re z > -\delta$, $\delta > 0$. Hence uniform modal observability cannot hold for the complete system. The conclusion follows since a finite-dimensional observable system is detectable. \square

The last result in particular implies that, for the control system described by (1), (2), modal observability is forced to coincide with detectability when (RE) admits a positive solution P . It is interesting to again discuss Example 4 from this point

of view. As was stated, it can be represented in the form (1), (2), and it is modally observable. It is not detectable, however. This follows from [12], since the null solution to the difference equation $\xi(t) = \xi(t-1)$ is not exponentially stable. Consequently, the Riccati equation relative to the system in this example does not have a positive solution.

5. Concluding remarks. In this paper we presented an analysis of the spectral properties of an operator A when an “algebraic” Riccati equation has a solution. The crucial properties that we assumed for the operator A and the operators H, C that appear in (RE) are that H is compact and that the pair (A, C) is modally detectable over a given region Ω . The case where H is nonpositive, $H = BB^* - DD^*$, in particular, is also considered. This case occurs in H^∞ -control theory.

In the case where $H = BB^*$, we were able to identify a class of systems for which modal observability and detectability are equivalent properties.

The result that we were looking for in this paper were suggested by a “closed-loop configuration” of the system. We refer to [39] for related “open-loop” results.

In this paper, the operators B, C in (1), (2) are bounded operators, so that boundary control problems (or problems with boundary observation) are not considered (we quote [4] for an extension of the results in [15] to the class of systems studied in [32]). The theory of the quadratic regulator for boundary control processes is now well developed (see the survey [18] for discussion examples and references). In the case where the input acts on the boundary, the term $\langle x, PHPx \rangle$ takes the form $\langle D^*A^*Px, D^*A^*Px \rangle$, where D solves a stationary “Dirichlet” problem. It is possible to prove that this operator is bounded when A generates an analytic semigroup and $B \in L(U, X_\alpha)$, where X_α is the domain of $(\lambda I - A)^\alpha$, $0 \leq \alpha \leq 1$. Here λ is any number such that $e^{-\lambda t}E(t)$ is exponentially stable (see [18] and references therein, in particular, [7], [10], [16], [17] listed there, for an analysis of boundary control parabolic systems). The results of this paper could be applied should the operator D^*A^*P be compact. However, this observation is not of much use, since the conditions imposed on the operator A to satisfy this condition (of being the generator of an analytic semigroup, with compact resolvent) already give a clear picture of its spectrum.

Some of the results in §§3, 4 do not require that A is the generator of a C_0 -semigroup; this is important in the analysis of “ill-posed” control problems. The “ill-posed” control problems studied until now are examples of inverse control problems, i.e., control problems described as in (1) (2), with $-A$ (and not A) being a generator of a C_0 -semigroup. It is clear that if P is a solution to the Riccati equation that corresponds to $-A$, then $-P$ is a solution to (RE). This observation is used in §4 to explain one of the reasons that force the solutions to the Riccati equation for inverse systems to be unbounded.

The final results in §4 shows that, for a control process described by (1), (2), the existence of a positive solution to (RE) forces modal observability to be equivalent to detectability. This is a direct extension of the results in [15].

Acknowledgments. The author thanks the referees, whose observations suggested that we present more details on the notion of modal observability. Moreover, they suggested a shorter proof for Theorem 15. The author also acknowledges comments received by R. Triggiani about the case of boundary control systems.

REFERENCES

- [1] W. ARENDT, A. GABOSCH, G. GREINER, U. GROH, H. P. LOTZ, U. MOUSTAKAS, R. NAGEL, F. NEUBANDER, AND U. SCHLOTTERBECK, *One-Parameter Semigroups of Positive Operators*, Lecture Notes in Mathematics, 1184, R. Nagel, ed., Springer-Verlag, Berlin, 1986.
- [2] K. P. M. BAHT AND W. M. WONHAM, *Stabilizability and detectability for evolution systems in Banach spaces*, in Proc. 1976 Conference on Decision and Control, IEEE Publications, New York, 1976, pp. 1240–1243.
- [3] F. CALLIER AND C. A. DESOER, *Multivariable Feedback Systems*, Springer-Verlag, New York, 1982.
- [4] R. CURTAIN, *Equivalence of input-output stability and exponential stability for infinite-dimensional systems*, Math. Systems Theory, 21 (1988), pp. 19–48.
- [5] R. DATKO, *An extension of a theorem of A. M. Lyapunov to semigroups of operators*, J. Math. Anal. Appl., 24 (1968), pp. 290–295.
- [6] ———, *Extending a theorem of Liapunov to Hilbert spaces*, J. Math. Analysis Appl., 32 (1970), pp. 610–616.
- [7] F. FLANDOLI, *Riccati equation arising in a boundary control problem with distributed parameters*, SIAM J. Control Optim., 22 (1984), pp. 76–86.
- [8] ———, *On the optimal control of non well posed systems with boundary control*, in Proc. Conference Distributed Parameter Systems, F. Kappel, F. Kunish, W. Schappacher, eds., Springer-Verlag, Berlin, 1985, pp. 179–191.
- [9] ———, *Dynamic programming approach to the optimal control of systems governed by non well posed Cauchy problems in Hilbert spaces*, Boll. Un. Mat. Ital. 6, 5-B (1986), pp. 177–195.
- [10] ———, *Algebraic Riccati equation arising in boundary control problems*, SIAM J. Control Optim., 25 (1987), pp. 612–636.
- [11] M. L. J. HAUTUS, *Controllability and observability conditions of linear autonomous systems*, Nederl. Akad. Wetensch. Proc. Ser. A, 72 (1969), pp. 443–448.
- [12] D. HENRY, *Linear autonomous neutral functional differential equations*, J. Differential Equations, 15 (1974), pp. 106–128.
- [13] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
- [14] V. KUČERA, *Discrete Linear Control, The Polynomial Equation Approach*, John Wiley, Chichester, UK, 1979.
- [15] C. A. JACOBSON AND C. N. NETT, *Linear state-space systems in infinite-dimensional space: the role and characterization of joint stabilizability/detectability*, IEEE Trans. Automat. Control, AC-33 (1988), pp. 541–549.
- [16] I. LASIECKA, *Unified theory for abstract parabolic boundary value problems—a semigroup approach*, Appl. Math. Optim., 6 (1980), pp. 287–334.
- [17] I. LASIECKA AND R. TRIGGIANI, *The regulator problem for parabolic equations with Diriclet boundary control*, Appl. Math. Optim., 16 (1987), pp. 147–168.
- [18] ———, *Algebraic Riccati equations arising in boundary/point control: a survey of theoretical and numerical results. Part I: continuous case*, in Perspective in Control Theory, B. Jacobczyk, K. Malanowski, and W. Respondek, eds., Birkhauser, Boston, 1990, pp. 175–210.
- [19] J. L. LIONS, *Contrôle des systèmes distribués singuliers*, Gauthier Villars, Paris, 1983.
- [20] H. LOGEMANN, *On the transfer matrix of a neutral system: characterization of exponential stability in input-output terms*, Systems Control Lett., 9 (1987), pp. 393–400.
- [21] J. C. LOUIS AND D. WEXLER, *Stability in Hilbert spaces using Riccati equation*, Equadiff 82, Proc. Intern. Conference, Würzburg Germany, 1982, Lecture Notes in Mathematics, Springer-Verlag, Berlin, 1983, pp. 440–445.
- [22] A. MANITIUS, *Necessary and sufficient conditions of approximate controllability for general linear retarded systems*, SIAM J. Control Optim., 19 (1981), pp. 516–532.
- [23] ———, *F-controllability and observability of linear retarded systems*, Appl. Math. Optim. 9 (1982), pp. 73–95.
- [24] A. W. OLBROT, *Stabilizability, detectability and spectrum assignment for linear systems with general time delays*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 887–890.
- [25] L. PANDOLFI, *Controllability properties of perturbed distributed parameter systems*, Linear Algebra Appl., 122/123/124 (1989), pp. 525–538.
- [26] ———, *Some properties of the frequency domain description of boundary control systems*, J. Math. Anal. Appl., 142 (1989), pp. 219–241.

- [27] L. PANDOLFI, *Generalized control systems, boundary control systems and delayed control systems*, Math. Control Signals Systems, 3 (1990), pp. 165–181.
- [28] ———, *Modal observability of boundary control systems*, Boll. Un. Mat. Ital., (7) 4–B (1990), pp. 285–294.
- [29] I. R. PETERSEN, *Disturbance attenuation and H^∞ optimization: a design method based on the algebraic Riccati equation*, IEEE Trans. Autom. Control AC-32 (1987), pp. 427–429.
- [30] P. H. RIVERA AND C. F. VASCONCELOS, *Optimal control for a backward parabolic system*, SIAM J. Control Optim., 25 (1987), pp. 1163–1172.
- [31] H. H. ROSENBROCK, *State-Space and Multivariable Theory*, Nelson, London, 1970.
- [32] D. SALAMON, *Infinite-dimensional linear systems with unbounded control and observations: a functional analytic approach*, Trans. Amer. Math. Soc., 300 (1987), pp. 383–431.
- [33] R. TRIGGIANI, *On the stabilizability problem in Banach spaces*, J. Math. Anal. Appl., 52 (1975), pp. 383–403.
- [34] ———, *On the lack of exact controllability for mild solutions in Banach spaces*, J. Math. Anal. Appl., 50 (1975), pp. 438–446.
- [35] G. WEISS, *Weak L^p -stability of a linear semigroup on a Hilbert space implies exponential stability*, J. Differential Equations, 76 (1988), pp. 269–285.
- [36] J. ZABCZYK, *A note on C_0 -semigroups*, Bull. Acad. Polon. Sci., 23 (1975), pp. 895–898.
- [37] ———, *Remarks on the algebraic Riccati equation in Hilbert space*, Appl. Math. Optim., 2 (1976), pp. 251–256.
- [38] ———, *Complete stabilizability implies exact controllability*, Seminarul de Ecuatii Functionale, University din Timisoara, Facultatea de Stiinte ale Naturii, Rumania, 1976.
- [39] H. ZWART, *Some remarks on open and closed loop stabilizability for infinite dimensional systems*, in Proc. Fourth Conf. on Distributed Parameter Systems, Vorau, Styria, Austria, 1988.

ON THE CONTINUOUS DEPENDENCE WITH RESPECT TO SAMPLING OF THE LINEAR QUADRATIC REGULATOR PROBLEM FOR DISTRIBUTED PARAMETER SYSTEMS*

I. G. ROSEN[†] AND C. WANG[‡]

Abstract. The convergence of solutions to the discrete- or sampled-time linear quadratic regulator problem and associated Riccati equation for infinite-dimensional systems to the solutions to the corresponding continuous time problem and equation, as the length of the sampling interval (the sampling rate) tends toward zero (infinity) is established. Both the finite- and infinite-time horizon problems are studied. In the finite-time horizon case, strong continuity of the operators that define the control system and performance index, together with a stability and consistency condition on the sampling scheme are required. For the infinite-time horizon problem, in addition, the sampled systems must be stabilizable and detectable, uniformly with respect to the sampling rate. Classes of systems for which this condition can be verified are discussed. Results of numerical studies involving the control of a heat/diffusion equation, a hereditary or delay system, and a flexible beam are presented and discussed.

Key words. LQR problem, feedback control, sampled control systems, approximation theory

AMS(MOS) subject classifications. 41A34, 49B27, 49B34, 65J10, 65J10

1. Introduction. In this paper we consider the convergence of closed-loop solutions to discrete- or sampled-time linear quadratic (LQ) optimal control problems and the associated Riccati equations for infinite-dimensional systems defined on Hilbert spaces to the solutions to the corresponding continuous-time problems and Riccati equations, as the length of the sampling interval tends toward zero. With the advent and proliferation of microcomputers, and control tasks becoming ever more complex (for example, the stabilization of large flexible spacecraft and fluid flow control), the roles played by discrete- or sampled-time control design techniques and distributed parameter systems have become increasingly more important. It has become necessary, therefore, to develop extensions of many of the familiar results for finite-dimensional systems to an infinite-dimensional setting. One area that has recently received a great deal of attention has been the LQ theory. Certain aspects of the linear-quadratic approach to control design for both continuous and sampled time infinite-dimensional systems have been studied extensively. In particular, these aspects include, for example, the linear state-feedback structure of the optimal control law, the optimal linear quadratic Gaussian (LQG) estimator and compensator problems, boundary control, and finite-dimensional approximation (for specific references, see below). To the best of our knowledge, however, the interrelation between the continuous- and discrete-time theories, which in the finite-dimensional case is well understood, has not as of yet been looked at in the context of infinite-dimensional systems. Such a study

* Received by the editors January 16, 1990; accepted for publication (in revised form) April 4, 1991.

[†] Center for Applied Mathematical Sciences, Department of Mathematics, University of Southern California, Los Angeles, California 90089-1113. This author was supported in part by grant AFOSR-87-0356 from the United States Air Force Office of Scientific Research. A portion of this research was carried out while the author was a visiting scientist at the Institute for Computer Applications in Science and Engineering (ICASE) at National Aeronautics and Space Administration (NASA) Langley Research Center in Hampton, Virginia, which is operated under NASA contract NAS1-18107.

[‡] Center for Applied Mathematical Sciences, Department of Mathematics, University of Southern California, Los Angeles, California 90089-1113. This author was supported in part by a grant from the University of Southern California Faculty Research and Innovation Fund (FRIF).

would be useful, for example, because typically in engineering practice, the discrete- and continuous-time LQ theories are applied interchangeably without regard to as to whether the actual system is continuous or discrete in nature. In particular, due to hardware constraints, most systems occurring in engineering practice are, in fact, discrete. However, if the sampling is considered to be rapid enough, the system may be treated as continuous when an optimal control law, state estimator, or compensator is designed. Our work is largely motivated by the fact that the results we present here will serve to, in some sense, justify this approach.

We note that in finite dimensions, where strong and uniform norm convergence of linear operators are equivalent, the continuous dependence with respect to sampling of the solution to the linear quadratic control problem and associated Riccati equation is straightforward. Indeed, in [Le], the continuous-time theory (for the LQG estimator) is established by first deriving the discrete-time results, which are fundamentally algebraic in nature, and then taking the limit as the length of the sampling interval tends toward zero. However, in infinite dimensions, as is typically the case, the desired convergence is less obvious. This is because we must deal with the convergence of operators given only as the solutions to nonlinear Riccati-type integral equations. The difficulties become especially acute in the case of the infinite-time horizon problem.

In this paper, we consider both the finite- and infinite-time horizon problems. In the case of the finite-time horizon problem, under the assumption of strong continuity of the operators that define the control system and performance index, together with a stability and consistency hypothesis on the sampling scheme, we are able to deduce the desired convergence. In doing this, we must develop an appropriate framework to facilitate the comparison of discrete- and continuous-time operator families. For this purpose, we rely heavily upon Kato's [K] treatment of discrete semigroups. In the case of the infinite-time horizon problem, we must additionally assume stabilizability and detectability of the discrete-time systems with some degree of uniformity in the sampling rate. The notion of stabilizability/detectability uniform with respect to sampling will be made precise in §3. We are able to establish that if the continuous-time system is stabilizable and detectable via finite rank feedback, and if zero-order hold sampling is employed, then the resulting discrete-time systems are uniformly stabilizable and detectable for sufficiently small sampling interval. We also have a result concerning the uniform stabilizability and detectability of systems, which are described open-loop by compact, analytic, or differentiable semigroups. However, this result is not discussed here, but rather in [RW] and [RW2].

Our treatment is functional analytic in nature, and is similar in spirit to the many recent studies of convergence of solutions to LQ control and estimation problems and the associated Riccati equations under state (space) approximation (i.e., finite difference, modal, or finite element, for example). See, for example, [BK], [BW], [G], [GA], [GR], [R], and [W]. For the discrete-time LQ theory for infinite-dimensional systems, we rely heavily on the well-known results contained in [HH], [LCB], and [Z].

In addition to our theoretical results, we have included the results of some of our numerical convergence studies. We present and discuss our findings for the infinite-time horizon LQ optimal control problems for a one-dimensional heat or diffusion equation, a one-dimensional hereditary or delay system, and a hybrid system of ordinary and partial differential equations describing the small amplitude transverse vibration of a cantilevered Voigt-Kelvin viscoelastic beam with tip mass.

An outline of the remainder of the paper is as follows. In §2 we treat the finite-time horizon problem. The infinite-time horizon problem is considered in §3. The

relation between finite rank and uniform stabilizability and detectability is treated in §4. Our numerical results are presented and discussed in §5, while a brief §6 contains a summary and some concluding remarks.

2. LQR problems with finite-time horizon. In this section we consider the linear quadratic regulator (LQR) problem over a finite-time interval. The basic notation and our general assumptions are introduced in the statements of both the continuous-time and corresponding sampled-time problems given below. The existence and uniqueness of the optimal control, as well as its closed-loop feedback structure, can be obtained using a variety of approaches. Here we consider the optimal control problem as the minimization of a strictly coercive quadratic form on the admissible control space. This approach yields an explicit representation for the solution of the usual Riccati equations (for both the continuous- and sampled-time problems) in terms of the underlying system and penalty operators that define the problems. Since the particular focus of our effort here is the consideration of sampled-time problems as approximations to a continuous-time problem, specialized notions and characterizations of convergence must be introduced. Once this is done, our convergence result for the finite-time horizon problem can then be stated in terms of these specialized notions of convergence. In the discussion to follow, Theorems 2.1 and 2.2 state the well-known existence and uniqueness results for the solutions to the continuous- and discrete-time LQR problems on a finite time-interval in closed-loop, linear state feedback form, while Theorem 2.3 is concerned with the specialized notions of convergence mentioned above. Our convergence results for the finite-time horizon problem is given in Theorem 2.4 and Corollary 2.1

Let H and U be Hilbert spaces with inner products $\langle \cdot, \cdot \rangle_H$ and $\langle \cdot, \cdot \rangle_U$, respectively. Let $t_0, t_f \in \mathbb{R}$ be given with $t_0 < t_f$, and let $T = \{T(t, s) : t_0 \leq s \leq t \leq t_f\}$ be an evolution system on H . For each $t \in [t_0, t_f]$, let $B(t) \in L(U, H)$, $Q(t) \in L(H)$, and $R(t) \in L(U)$, and let $G \in L(H)$. We consider the continuous-time LQR problem given by the following problem.

(P) Determine a control input $\bar{u} \in L_2(t_0, t_f; U)$ which minimizes the quadratic performance index

$$J(u; t_0, x(t_0), G) = \langle Gx(t_f), x(t_f) \rangle_H + \int_{t_0}^{t_f} \{ \langle Q(t)x(t), x(t) \rangle_H + \langle R(t)u(t), u(t) \rangle_U \} dt,$$

where for each $t \in [t_0, t_f]$ the state $x(t) \in H$ is given by

$$(2.1) \quad x(t) = T(t, s)x(s) + \int_s^t T(t, \tau)B(\tau)u(\tau)d\tau, \quad t_0 \leq s \leq t \leq t_f.$$

We make the following standard assumptions on the operator families $\{T, B, G, Q, R\}$, which determine problem (P).

(C1) The evolution system T is strongly continuous on H and, therefore, is uniformly exponentially bounded, with constants $M > 0$ and $\omega \in \mathbb{R}$. That is,

$$\|T(t, s)\|_{L(H)} \leq Me^{\omega(t-s)}, \quad t_0 \leq s \leq t \leq t_f.$$

(C2) The operator-valued functions B, Q , and R are strongly continuous and, therefore, are uniformly bounded on $[t_0, t_f]$. That is, there exists a constant $C > 0$ for which

$$\|B(t)\|_{L(U, H)} \leq C, \quad \|Q(t)\|_{L(H)} \leq C, \quad \|R(t)\|_{L(U)} \leq C,$$

$t \in [t_0, t_f]$.

(C3) The operator G and the operators $Q(t)$ and $R(t)$ for each $t \in [t_0, t_f]$ are selfadjoint and nonnegative definite. Moreover, there exists a constant $r > 0$ for which $R(t) \geq rI$, $t \in [t_0, t_f]$.

The strong continuity assumption in (C2) is not necessary for the wellposedness of the LQR problem. However, some assumptions on the continuity of the operators B, Q, R will be needed to obtain uniform convergence with respect to sampling.

The closed-loop linear state-feedback form of the solution to problem (P) can be shown to exist and can be explicitly constructed by considering the minimization of appropriately constructed, strictly coercive, quadratic forms on the Hilbert spaces $\mathcal{U}_s = L_2(s, t_f; U)$, $s \in [t_0, t_f]$ (see, for example, [G]). For each $s \in [t_0, t_f]$ define the operators $\mathcal{B}_s \in L(H, \mathcal{U}_s)$ and $\mathcal{R}_s \in L(\mathcal{U}_s)$ by

$$(2.2) \quad (\mathcal{B}_s \phi)(t) = B(t)^* \left\{ T(t_f, t)^* G T(t_f, s) + \int_t^{t_f} T(\eta, t)^* Q(\eta) T(\eta, s) d\eta \right\} \phi,$$

for $\phi \in H$, $t \in [s, t_f]$, and

$$(2.3) \quad (\mathcal{R}_s u_s)(t) = R(t) u_s(t) + B(t)^* T(t_f, t)^* G \int_s^{t_f} T(t_f, \eta) B(\eta) u_s(\eta) d\eta \\ + B(t)^* \int_t^{t_f} \left\{ T(\eta, t)^* Q(\eta) \int_s^\eta T(\eta, \tau) B(\tau) u_s(\tau) d\tau \right\} d\eta,$$

for $t \in [s, t_f]$ and $u_s \in \mathcal{U}_s$. The adjoint operator $\mathcal{B}_s^* \in L(\mathcal{U}_s, H)$ of \mathcal{B}_s is given by

$$(2.4) \quad \mathcal{B}_s^* u_s = T(t_f, s)^* G \int_s^{t_f} T(t_f, t) B(t) u_s(t) dt \\ + \int_s^{t_f} T(\tau, s)^* Q(\tau) \left\{ \int_s^\tau T(\tau, \eta) B(\eta) u_s(\eta) d\eta \right\} d\tau.$$

For $x(s) \in H$ given, $J(\cdot; s, x(s), G)$ is minimized by choosing $u_s = \bar{u}_s = -\mathcal{R}_s^{-1} \mathcal{B}_s x(s) \in \mathcal{U}_s$. It can be shown that (see [G], [RW1])

$$\min_{\mathcal{U}_s} J(\cdot; s, x(s), G) = J(\bar{u}_s; s, x(s), G) \\ = \langle G T(t_f, s) x(s), T(t_f, s) x(s) \rangle_H + \int_s^{t_f} \langle Q(t) T(t, s) x(s), T(t, s) x(s) \rangle_H dt \\ - \langle \mathcal{R}_s^{-1} \mathcal{B}_s x(s), \mathcal{B}_s x(s) \rangle_{\mathcal{U}_s} \\ = \langle \Pi(s) x(s), x(s) \rangle_H,$$

where the selfadjoint operator-valued function $\Pi : [t_0, t_f] \mapsto L(H)$ is defined by

$$(2.5) \quad \Pi(s) \phi = T(t_f, s)^* G T(t_f, s) \phi + \int_s^{t_f} T(t, s)^* Q(t) T(t, s) \phi dt \\ - \mathcal{B}_s^* \mathcal{R}_s^{-1} \mathcal{B}_s \phi, \quad \phi \in H.$$

Using the definitions given above, the following theorem concerning the existence and characterization of the closed-loop solution to problem (P) can be established.

THEOREM 2.1. *Suppose that assumptions (C1)–(C3) are satisfied. Then for any initial state $x(t_0) \in H$ given, there exists a unique solution \bar{u} to problem (P). The optimal control \bar{u} is given in linear state-feedback form by*

$$\bar{u}(t) = -R(t)^{-1} B(t)^* \Pi(t) \bar{x}(t), \quad t \in [t_0, t_f],$$

where \bar{x} is the optimal trajectory. The operator-valued function Π is given by (2.5) and it is the unique selfadjoint solution to the Riccati integral equation

$$(2.6) \quad \Pi(t) = T(t_f, t)^* G T(t_f, t) + \int_t^{t_f} T(\tau, t)^* \{Q(\tau) - \Pi(\tau) B(\tau) R(\tau)^{-1} B(\tau)^* \Pi(\tau)\} T(\tau, t) d\tau,$$

$t \in [t_0, t_f]$. We have

$$(2.7) \quad \min_{\mathcal{U}_s} J(\cdot; t_0, x(t_0), G) = J(\bar{u}; t_0, x(t_0), G) = \langle \Pi(t_0)x(t_0), x(t_0) \rangle_H.$$

We consider next the discrete- or sampled-time problem. Let $k_0, k_f \in Z$ with $k_f > k_0$ and let $h \in R$ with $h > 0$. For $k \in Z$ with $k_0 \leq k \leq k_f - 1$ let $A_h(k) \in L(H)$, and let $\{T_h(k, j) : k_0 \leq j \leq k \leq k_f\}$ be the discrete-time evolution system on H given by

$$(2.8) \quad T_h(k, k) = I, \\ T_h(k, j) = A_h(k-1) \cdot A_h(k-2) \cdots A_h(j) = \prod_{i=j}^{k-1} A_h(i), \quad k_0 \leq j < k \leq k_f.$$

Let $\{B_h(k)\}_{k=k_0}^{k_f-1}$, $\{Q_h(k)\}_{k=k_0}^{k_f-1}$, and $\{R_h(k)\}_{k=k_0}^{k_f-1}$ be sequences in $L(U, H)$, $L(H)$, and $L(U)$, respectively, and let $G_h \in L(H)$. The LQR problem is then given by

(P_h) Determine a control input $\bar{u}_h \in l_2(k_0, k_f-1; U)$ that minimizes the quadratic performance index

$$J_h(u_h; k_0, x_h(k_0), G_h) = \langle G_h x_h(k_f), x_h(k_f) \rangle_H + h \sum_{k=k_0}^{k_f-1} \{ \langle Q_h(k)x_h(k), x_h(k) \rangle_H + \langle R_h(k)u_h(k), u_h(k) \rangle_U \},$$

where for each $k \in Z$ with $k_0 < k \leq k_f$, the state $x_h(k) \in H$ is given by

$$(2.9) \quad x_h(k) = T_h(k, j)x_h(j) + h \sum_{i=j}^{k-1} T_h(k, i+1)B_h(i)u_h(i),$$

for $k_0 \leq j < k \leq k_f$.

For the discrete-time case, we make the following assumptions.

(D1) For each $h > 0$ the operators $A_h(k)$, $B_h(k)$, $Q_h(k)$, and $R_h(k)$ are bounded in k for $k_0 \leq k \leq k_f - 1$. Thus, there exists a constant C_h for which

$$\begin{aligned} \|A_h(k)\|_{L(H)} &\leq C_h, & \|B_h(k)\|_{L(U, H)} &\leq C_h, \\ \|Q_h(k)\|_{L(H)} &\leq C_h, & \|R_h(k)\|_{L(U)} &\leq C_h, \end{aligned}$$

for $k_0 \leq k \leq k_f - 1$.

(D2) The operator G_h and the operators $Q_h(k)$ and $R_h(k)$ for $k_0 \leq k \leq k_f - 1$ are selfadjoint and nonnegative. Moreover, there exists a constant $r_h > 0$ for which $R_h(k) \geq r_h I$, $k_0 \leq k \leq k_f - 1$.

Note that assumption (D1), together with (2.8), yields that the discrete-time evolution system $\{T_h(k, j) : k_0 \leq j \leq k \leq k_f\}$ is uniformly exponentially bounded with

$$\|T_h(k, j)\|_{L(H)} \leq C_h^{k-j}, \quad k_0 \leq j \leq k \leq k_f.$$

Note also that the discrete-time evolution equation (2.9) is equivalent to the discrete-time dynamical system given by

$$(2.10) \quad x_h(k+1) = A_h(k)x_h(k) + hB_h(k)u_h(k), \quad k_0 \leq k \leq k_f - 1, x_h(k_0) \in H.$$

For each $h > 0$ and $j = k_0, k_0 + 1, \dots, k_f - 1$, let $\mathcal{U}_{h,j} = l_2(j, k_f - 1; U)$ endowed with the inner product

$$\langle u_{h,j}, v_{h,j} \rangle_{\mathcal{U}_{h,j}} = h \sum_{k=j}^{k_f-1} \langle u_{h,j}(k), v_{h,j}(k) \rangle_U.$$

Define the operators $\mathcal{B}_{h,j} \in L(H, \mathcal{U}_{h,j})$ and $\mathcal{R}_{h,j} \in L(\mathcal{U}_{h,j})$ by

$$(2.11) \quad (\mathcal{B}_{h,j}\phi)(k) = B_h(k)^*T_h(k_f, k+1)^*G_hT_h(k_f, j)\phi \\ + B_h(k)^* \left\{ h \sum_{i=k+1}^{k_f-1} T_h(i, k+1)^*Q_h(i)T_h(i, j) \right\} \phi,$$

for $\phi \in H$, $k = j, j+1, \dots, k_f - 1$, and

$$(2.12) \quad (\mathcal{R}_{h,j}u_{h,j})(k) = R_h(k)u_{h,j}(k) \\ + B_h(k)^*T_h(k_f, k+1)^*G_hh \sum_{i=j}^{k_f-1} T_h(k_f, i+1)B_h(i)u_{h,j}(i) \\ + B_h(k)^*h \sum_{i=k+1}^{k_f-1} T_h(i, k+1)^*Q_h(i) \left\{ h \sum_{l=j}^{i-1} T_h(i, l+1)B_h(l)u_{h,j}(l) \right\},$$

$u_{h,j} \in \mathcal{U}_{h,j}$, $k = j, j+1, \dots, k_f - 1$, respectively, where in the above expressions and throughout the remainder of the paper, we adopt the convention that $\sum_{i=\mu}^{\nu} a_i = 0$ whenever $\nu < \mu$. The adjoint of $\mathcal{B}_{h,j}$, the operator $\mathcal{B}_{h,j}^* \in L(\mathcal{U}_{h,j}, H)$ is given by

$$(2.13) \quad \mathcal{B}_{h,j}^*u_{h,j} = T_h(k_f, j)^*G_hh \sum_{k=j}^{k_f-1} T_h(k_f, k+1)B_h(k)u_{h,j}(k) \\ + h \sum_{k=j+1}^{k_f-1} T_h(k, j)^*Q_h(k) \left\{ h \sum_{i=j}^{k-1} T_h(k, i+1)B_h(i)u_{h,j}(i) \right\},$$

for $u_{h,j} \in \mathcal{U}_{h,j}$.

As in the continuous-time case, for $j = k_0, \dots, k_f - 1$, $x_h(j) \in H$, and $u_{h,j} \in \mathcal{U}_{h,j}$

$$J_h(u_{h,j}; j, x_h(j), G_h) = \langle G_hT_h(k_f, j)x_h(j), T_h(k_f, j)x_h(j) \rangle_H \\ + h \sum_{k=j}^{k_f-1} \langle Q_h(k)T_h(k, j)x_h(j), T_h(k, j)x_h(j) \rangle_H \\ - \langle \mathcal{R}_{h,j}^{-1}\mathcal{B}_{h,j}x_h(j), \mathcal{B}_{h,j}x_h(j) \rangle_{\mathcal{U}_{h,j}} \\ + \langle \mathcal{R}_{h,j}(u_{h,j} + \mathcal{R}_{h,j}^{-1}\mathcal{B}_{h,j}x_h(j)), u_{h,j} + \mathcal{R}_{h,j}^{-1}\mathcal{B}_{h,j}x_h(j) \rangle_{\mathcal{U}_{h,j}},$$

where the existence of the inverse of $\mathcal{R}_{h,j}$ is guaranteed by assumption (D2). For $j \in Z$ with $j \in [k_0, k_f - 1]$ and $x_h(j) \in H$ given, $J_h(\cdot; j, x_h(j), G_h)$ is minimized when $u_{h,j} = \bar{u}_{h,j} = -\mathcal{R}_{h,j}^{-1} \mathcal{B}_{h,j} x_h(j)$;

$$\begin{aligned} \min_{\mathcal{U}_{h,j}} J_h(\cdot; j, x_h(j), G_h) &= J_h(\bar{u}_{h,j}; j, x_h(j), G_h) \\ &= \langle G_h T_h(k_f, j) x_h(j), T_h(k_f, j) x_h(j) \rangle_H \\ &\quad + h \sum_{k=j}^{k_f-1} \langle Q_h(k) T_h(k, j) x_h(j), T_h(k, j) x_h(j) \rangle_H \\ &\quad - \langle \mathcal{R}_{h,j}^{-1} \mathcal{B}_{h,j} x_h(j), \mathcal{B}_{h,j} x_h(j) \rangle_{\mathcal{U}_{h,j}} \\ &= \langle \Pi_h(j) x_h(j), x_h(j) \rangle_H, \end{aligned}$$

where the sequence of selfadjoint operators in $L(H)$, $\{\Pi_h(k)\}_{k=k_0}^{k_f-1}$, are given by

$$\begin{aligned} (2.14) \quad \Pi_h(j) \phi &= T_h(k_f, j)^* G_h T_h(k_f, j) \phi + h \sum_{k=j}^{k_f-1} T_h(k, j)^* Q_h(k) T_h(k, j) \phi \\ &\quad - \mathcal{B}_{h,j}^* \mathcal{R}_{h,j}^{-1} \mathcal{B}_{h,j} \phi, \end{aligned}$$

for $k = k_0, \dots, k_f - 1$ and $\phi \in H$. It is completely consistent to define $\Pi_h(k_f) = G_h$.

Using the above definitions, it is then possible to establish the following well-known result (see, for example, [LCB], [Z], and [GR]) for the discrete-time LQR problem (P_h).

THEOREM 2.2. *Suppose that assumptions (D1) and (D2) are satisfied. Then for any given initial state $x_h(k_0) \in H$ there exists a unique solution $\bar{u}_h \in l_2(k_0, k_f - 1; U)$ to problem (P_h). It is given in linear state-feedback form by*

$$\bar{u}_h(k) = -\hat{R}_h(k)^{-1} B_h(k)^* \Pi_h(k+1) A_h(k) \bar{x}_h(k), \quad k = k_0, \dots, k_f - 1,$$

where $\hat{R}_h(k) = R_h(k) + h B_h(k)^* \Pi_h(k+1) B_h(k)$, for $k = k_0, \dots, k_f - 1$, and the optimal trajectory \bar{x}_h is given by (2.9) (equivalently, (2.10)) with $u_h = \bar{u}_h$. The sequence of operators in $L(H)$, $\{\Pi_h(k)\}_{k=k_0}^{k_f-1}$ is given by (2.14) with $\Pi_h(k_f) = G_h$ and can be obtained recursively via the Riccati difference equation

$$\begin{aligned} (2.15) \quad \Pi_h(k) &= A_h(k)^* \Pi_h(k+1) A_h(k) \\ &\quad - h A_h(k)^* \Pi_h(k+1) B_h(k) \hat{R}_h(k)^{-1} B_h(k)^* \Pi_h(k+1) A_h(k) \\ &\quad + h Q_h(k), \end{aligned}$$

$k = k_f - 1, \dots, k_0$, $\Pi_h(k_f) = G_h$. We have

$$\begin{aligned} (2.16) \quad \min_{\mathcal{U}_{h,k_0}} J_h(\cdot; k_0, x_h(k_0), G_h) &= J_h(\bar{u}_h; k_0, x_h(k_0), G_h) \\ &= \langle \Pi_h(k_0) x_h(k_0), x_h(k_0) \rangle_H. \end{aligned}$$

For appropriate choices of the families of operators T_h , B_h , Q_h , and R_h , we are interested in studying the convergence of solutions to the problems (P_h) to the solution of problem (P) as the length of the sampling interval h tends toward zero. In particular, we want to investigate the convergence of the discrete families of Riccati operators $\{\Pi_h(k) : k_0 \leq k \leq k_f\}$ to the continuous family of operators $\{\Pi(t) : t_0 \leq t \leq t_f\}$.

To reduce the necessary degree of technical detail, we make the simplifying assumption that $t_0 = 0$. There is, of course, no loss of generality in doing this since any system can be transformed to one on a time interval starting at the origin. Set $k_0 = 0$ and for each $h > 0$ let $k_f = k_{f,h} = [t_f/h]$ where for $a \in R$, $[a]$ is used to denote the greatest integer less than or equal to a . Let $t_{f,h} = hk_{f,h}$ and note that $\lim_{h \rightarrow 0^+} t_{f,h} = t_f$.

To compare discrete and continuous families of operators, it is useful to identify certain l_2 sequence spaces with subspaces of L_2 . For X a Hilbert space and all $h > 0$, let $L_{2,h}(0, t_{f,h}; X)$ be the subspace of $L_2(0, t_{f,h}; X)$ defined by $L_{2,h}(0, t_{f,h}; X) = \{\phi \in L_2(0, t_{f,h}; X) : \phi \text{ is constant on each of the intervals } [0, h), [h, 2h), \dots, [(k_{f,h} - 1)h, t_{f,h})\}$. Note that the subspace $L_{2,h}(0, t_{f,h}; X)$ of $L_2(0, t_{f,h}; X)$ is isometrically isomorphic to the space $l_2(0, k_{f,h} - 1; X)$ endowed with the inner product

$$\langle \{\phi_j\}_{j=0}^{k_{f,h}-1}, \{\psi_j\}_{j=0}^{k_{f,h}-1} \rangle = h \sum_{j=0}^{k_{f,h}-1} \langle \phi_j, \psi_j \rangle_X.$$

Let $\mathcal{U} = L_2(0, t_f; U)$ and let $\mathcal{U}_h = L_{2,h}(0, t_{f,h}; U)$. Let $P_h \in L(\mathcal{U}, \mathcal{U}_h)$ be the orthogonal projection-like mapping of \mathcal{U} onto \mathcal{U}_h defined by

$$(P_h \phi)(t) = \sum_{j=0}^{k_{f,h}-1} (\phi_h)_j \chi_{I_j}(t), \quad 0 \leq t \leq t_{f,h},$$

for $\phi \in \mathcal{U}$ where for $j = 0, 1, \dots, k_{f,h} - 1$, χ_{I_j} is the characteristic function for the interval $I_j = [jh, (j+1)h)$ and

$$(\phi_h)_j = h^{-1} \int_{I_j} \phi(t) dt.$$

It is not difficult to show (see [RW1]) that

1. the net $\{\|P_h\|_{L(\mathcal{U}, \mathcal{U}_h)}\}_{h>0}$ is uniformly bounded;
2. $\lim_{h \rightarrow 0^+} \|P_h \phi\|_{\mathcal{U}_h} = \|\phi\|_{\mathcal{U}}$, $\phi \in \mathcal{U}$, and
3. for each $\psi \in \mathcal{U}_h$ there exists a $\phi \in \mathcal{U}$ such that $\psi = P_h \phi$ and $\|\phi\|_{\mathcal{U}} = \|\psi\|_{\mathcal{U}_h}$.

Following Kato [K, §IX.4] we say that a net $\{\phi_h\}_{h>0}$, $\phi_h \in \mathcal{U}_h$ converges to $\phi \in \mathcal{U}$ ($\phi_h \rightarrow \phi$, or $\lim_{h \rightarrow 0^+} \phi_h = \phi$) if

$$\lim_{h \rightarrow 0^+} \|\phi_h - P_h \phi\|_{\mathcal{U}_h} = 0.$$

Also, if for $h > 0$, $\Phi_h \in L(\mathcal{U}_h)$, then we say that Φ_h converges strongly to $\Phi \in L(\mathcal{U})$ if $\Phi_h P_h \phi \rightarrow \Phi \phi$, $\phi \in \mathcal{U}$; that is, if

$$\lim_{h \rightarrow 0^+} \|\Phi_h P_h \phi - P_h \Phi \phi\|_{\mathcal{U}_h} = 0, \quad \phi \in \mathcal{U}.$$

With strong operator convergence defined in this way, it can be shown that $\Phi_h P_h \phi \rightarrow \Phi \phi$, $\phi \in \mathcal{U}$ implies that the net $\{\|\Phi_h\|_{L(\mathcal{U}_h)}\}$ is uniformly bounded and that if $\Phi_h P_h \phi \rightarrow \Phi \phi$, and $\Psi_h P_h \phi \rightarrow \Psi \phi$, $\phi \in \mathcal{U}$, then $\Phi_h \Psi_h P_h \phi \rightarrow \Phi \Psi \phi$, $\phi \in \mathcal{U}$, etc. We note, of course, that an analogous definition of strong convergence can be made for bounded operators having only one or the other of its domain and co-domain being \mathcal{U}_h . That is, for example, if $\Phi_h \in L(X, \mathcal{U}_h)$ and $\Phi \in L(X, \mathcal{U})$ where X is a normed linear space, then we say that Φ_h converges strongly to Φ if $\Phi_h x \rightarrow \Phi x$, $x \in X$, or

$$\lim_{h \rightarrow 0^+} \|\Phi_h x - P_h \Phi x\|_{\mathcal{U}_h} = 0.$$

Following the treatment of discrete semigroups in Kato [K], we make the following formal definition.

DEFINITION 2.1. The discrete-time families of bounded linear operators $\Phi_h = \{\Phi_h(k_n, k_{n-1}, \dots, k_1) : 0 \leq k_1 \leq k_2 \leq \dots \leq k_n \leq K_h\}$, $h > 0$ from a Banach space X into a Banach space Y will be said to (strongly) approximate a continuous-time family of operators $\Phi = \{\Phi(t_n, t_{n-1}, \dots, t_1) : 0 \leq t_1 \leq t_2 \leq \dots \leq t_n \leq T\}$ with $\Phi(t_n, \dots, t_1) \in L(X, Y)$ for $t = (t_n, \dots, t_1) \in \Delta(n, T) = \{(t_n, t_{n-1}, \dots, t_1) \in R^n : 0 \leq t_1 \leq t_2 \leq \dots \leq t_n \leq T\}$, at $\hat{t} = (\hat{t}_n, \dots, \hat{t}_1) \in \Delta(n, T)$, if

1. There exists at least one net of multi-indices $\{\hat{k}_h\}_{h>0}$, $\hat{k} = (\hat{k}_{n,h}, \dots, \hat{k}_{1,h}) \in Z^n$ with $0 \leq \hat{k}_{1,h} \leq \dots \leq \hat{k}_{n,h} \leq K_h$ and $\lim_{h \rightarrow 0^+} h\hat{k}_h = \hat{t}$.
2. For all nets $\{k_h\}_{h>0}$, satisfying (i) above,

$$\lim_{h \rightarrow 0^+} \|\Phi_h(k_h)x - \Phi(\hat{t})x\|_Y = 0, \quad x \in X.$$

The families Φ_h , $h > 0$ will be said to approximate Φ on the set $\Delta(n, T)$, if $K_h = [T/h]$ and if Φ_h approximates Φ at each $\hat{t} \in \Delta(n, T)$.

When the discrete-time families Φ_h , $h > 0$ approximate the continuous-time family Φ at time \hat{t} (on the set $\Delta(n, T)$) we write $\Phi_h \rightarrow \Phi$ at time \hat{t} (on the set $\Delta(n, T)$).

DEFINITION 2.2. For $h > 0$ and $\Phi_h = \{\Phi_h(k_n, k_{n-1}, \dots, k_1) : 0 \leq k_1 \leq k_2 \leq \dots \leq k_n \leq K_h\}$ a discrete-time family of bounded linear operators, we define an associated continuous-time family of operators, $\tilde{\Phi}_h = \{\tilde{\Phi}_h(t_n, t_{n-1}, \dots, t_1) : 0 \leq t_1 < t_2 + h < \dots < t_n + h < (K_h + 1)h\}$ via $\tilde{\Phi}_h(t_n, \dots, t_1) = \Phi_h([t_n/h], \dots, [t_1/h])$ for $t = (t_n, \dots, t_1) \in \Delta_h(n, K_h) = \{(t_n, t_{n-1}, \dots, t_1) \in R^n : 0 \leq t_1 < t_2 + h < \dots < t_n + h < (K_h + 1)h\}$.

Note that when $K_h = [T/h]$, $\Delta(n, T) \subset \Delta_h(n, K_h)$ for all $h > 0$.

The proof of the following theorem can be argued in much the same manner as were the proofs of Lemmas IX.3.4 and IX.3.5 in Kato [K].

THEOREM 2.3. Suppose that the continuous-time family of bounded linear operators Φ is strongly continuous on $\Delta(n, T)$ and that Φ_h , $h > 0$ are discrete-time families for which $\Phi_h \rightarrow \Phi$ on the set $\Delta(n, T)$. Suppose further that for each $h > 0$, $\tilde{\Phi}_h$ is the continuous-time family on $\Delta_h(n, K_h)$ corresponding to the discrete-time family Φ_h constructed according to Definition 2.2 above. Then

- (i) The families Φ_h , $h > 0$ are uniformly bounded in h in $L(X, Y)$; that is there exists a constant $M > 0$ independent of h for which

$$\|\Phi_h(k_n, k_{n-1}, \dots, k_1)\|_{L(X, Y)} \leq M, \quad 0 \leq k_1 \leq k_2 \leq \dots \leq k_n \leq K_h, h > 0,$$

- (ii) $\tilde{\Phi}_h \rightarrow \Phi$ uniformly in t for $t \in \Delta(n, T)$; that is

$$\lim_{h \rightarrow 0^+} \|\tilde{\Phi}_h(t)x - \Phi(t)x\|_Y = 0, x \in X,$$

uniformly in t for $t = (t_n, \dots, t_1) \in \Delta(n, T)$.

Conversely, if $K_h = [T/h]$ and $\tilde{\Phi}_h \rightarrow \Phi$ uniformly in t for $t \in \Delta(n, T)$, then $\Phi_h \rightarrow \Phi$ on the set $\Delta(n, T)$.

Let the continuous-time families $T = \{T(t, s) : 0 \leq s \leq t \leq T\} \subset L(H)$, $B = \{B(t) : 0 \leq t \leq t_f\} \subset L(U, H)$, $Q = \{Q(t) : 0 \leq t \leq t_f\} \subset L(H)$ and $R = \{R(t) : 0 \leq t \leq t_f\} \subset L(U)$ be as given in the statement of the continuous-time LQR problem (P) (i.e., in particular assume that the conditions (C1)–(C3) hold). For $h > 0$, let $k_{f,h} = [t_f/h]$ and let $A_h = \{A_h(k) : 0 \leq k \leq k_{f,h} - 1\} \subset L(H)$, $B_h = \{B_h(k) : 0 \leq k \leq k_{f,h} - 1\} \subset L(U, H)$, $Q_h = \{Q_h(k) : 0 \leq k \leq k_{f,h} - 1\} \subset L(H)$, and

$R_h = \{R_h(k) : 0 \leq k \leq k_{f,h} - 1\} \subset L(U)$ be discrete time families of bounded linear operators, which satisfy conditions (D1) and (D2) and which satisfy the following conditions.

(A1) $B_h \rightarrow B$, $Q_h \rightarrow Q$, $R_h \rightarrow R$, and $B_h^* \rightarrow B^*$ on the set $\Delta(1, t_f)$ where $B^* = \{B(t)^* : 0 \leq t \leq t_f\}$ and $B_h^* = \{B_h(k)^* : 0 \leq k \leq k_{f,h}\}$.

(A2) (a) Stability. The discrete-time families of operators $T_h = \{T_h(k, j) : 0 \leq j \leq k \leq k_{f,h}\} \subset L(H)$ given by

$$T_h(k, j) = \begin{cases} \prod_{i=j}^{k-1} A_h(i), & j < k, \\ I, & j = k \end{cases}$$

are uniformly bounded in $L(H)$ for $h > 0$.

(b) Consistency.

$$\lim_{h \rightarrow 0^+} \frac{1}{h} \|\tilde{T}_h(t+h, t)\phi - T(t+h, t)\phi\| = 0, \quad \phi \in H,$$

and

$$\lim_{h \rightarrow 0^+} \frac{1}{h} \|\tilde{T}_h(t+h, t)^*\phi - T(t+h, t)^*\phi\| = 0, \quad \phi \in H,$$

uniformly in t for $t \in [0, t_f]$.

(A3) The scalars r_h given in the statement of condition (D2) are bounded away from zero uniformly in h . That is $r_h \geq r > 0$, $h > 0$.

LEMMA 2.1. Condition (A2) implies that $T_h \rightarrow T$ and $T_h^* \rightarrow T^*$ on the set $\Delta(2, t_f)$.

Proof. We consider the convergence $T_h \rightarrow T$ only; the adjoint convergence is completely analogous. Following the proof of the well-known Lax–Equivalence theorem [RM], the result is an immediate consequence of condition (A2), the strong continuity of the continuous-time family T , and the identity

$$T_h(k, j)\phi - T(kh, jh)\phi = \begin{cases} \sum_{i=j}^{k-1} T_h(k, i+1)\{A_h(i) - T((i+1)h, ih)\}T(ih, jh)\phi, & k > j, \\ 0, & k = j, \end{cases}$$

$0 \leq j \leq k \leq k_{f,h}$, $\phi \in H$. \square

We also assume that $G \in L(H)$ is as in condition (C3) and that for each $h > 0$ the operator $G_h \in L(H)$ satisfies condition (D2). We require that the additional approximation condition

(A4) $\lim_{h \rightarrow 0^+} G_h\phi = G\phi$, $\phi \in H$, be satisfied as well.

For $h > 0$ and $s \in [0, t_f]$ define $\tilde{B}_{h,s} \in L(H, \mathcal{U}_h)$ by

$$(2.17) \quad (\tilde{B}_{h,s}\phi)(t) = \begin{cases} \text{For } t \in [0, t_{f,h} - h] : \\ \chi_{[s/h]h, t_{f,h}}(t) \tilde{B}_h(t)^* \{\tilde{T}_h(t_{f,h}, t+h)^* G_h \tilde{T}_h(t_{f,h}, s) \\ + \int_{[(t+h)/h]h}^{t_{f,h}} \tilde{T}_h(\eta, t+h)^* \tilde{Q}_h(\eta) \tilde{T}_h(\eta, s) d\eta\} \phi, \\ \text{For } t \in [t_{f,h} - h, t_{f,h}] : \\ (\tilde{B}_{h,s}\phi)(t_{f,h} - h), \end{cases}$$

when $s \in [0, t_{f,h})$, and by $(\tilde{\mathcal{B}}_{h,s}) = 0$ when $s \in [t_{f,h}, t_f]$, for $\phi \in H$. Note that for $j = 0, 1, 2, \dots, k_{f,h} - 1$, $k = j, j + 1, \dots, k_{f,h} - 1$, and $\phi \in H$

$$(2.18) \quad (\tilde{\mathcal{B}}_{h,s}\phi)(t) = (\mathcal{B}_{h,j}\phi)(k),$$

for $s \in [jh, (j+1)h)$ and $t \in [kh, (k+1)h)$, where for $j = 0, 1, 2, \dots, k_{f,h} - 1$, $\mathcal{B}_{h,j} \in L(U, \mathcal{U}_{h,j})$ is given by (2.11).

For $s \in [0, t_f]$ the adjoint of the operator $\tilde{\mathcal{B}}_{h,s}$ given in (2.17), $\tilde{\mathcal{B}}_{h,s}^* \in L(\mathcal{U}_h, H)$, is given by

$$(2.19) \quad \begin{aligned} \tilde{\mathcal{B}}_{h,s}^* u_h &= \tilde{T}_h(t_{f,h}, s)^* G_h \int_{[s/h]h}^{t_{f,h}} \tilde{T}_h(t_{f,h}, t+h) \tilde{B}_h(t) u_h(t) dt \\ &+ \int_{[(s+h)/h]h}^{t_{f,h}} \tilde{T}_h(\tau, s)^* \tilde{Q}_h(\tau) \left\{ \int_{[s/h]h}^{[\tau/h]h} \tilde{T}_h(\tau, \eta+h) \tilde{B}_h(\eta) u_h(\eta) d\eta \right\} d\tau, \end{aligned}$$

when $s \in [0, t_{f,h})$, and by $\tilde{\mathcal{B}}_{h,s}^* u_h = 0$ when $s \in [t_{f,h}, t_f]$ for $u_h \in \mathcal{U}_h$. Note that for $j = 0, 1, 2, \dots, k_{f,h} - 1$ and $u_{h,j} \in \mathcal{U}_{h,j}$ we have

$$(2.20) \quad \tilde{\mathcal{B}}_{h,s}^* u_h = \mathcal{B}_{h,j}^* u_{h,j}$$

for $s \in [jh, (j+1)h)$, when $u_h \in \mathcal{U}_h$ is given by

$$(2.21) \quad u_h(t) = \begin{cases} 0, & 0 \leq t < jh, \\ u_{h,j}(k), & kh \leq t < (k+1)h, \end{cases}$$

$k = j, j + 1, \dots, k_{f,h} - 1$, and $\mathcal{B}_{h,j}^* \in L(\mathcal{U}_{h,j}, H)$ is given by (2.13).

For $s \in [0, t_f]$ define $\tilde{\mathcal{R}}_{h,s} \in L(\mathcal{U}_h)$ by

$$(2.22) \quad (\tilde{\mathcal{R}}_{h,s} u_h)(t) = \begin{cases} \text{For } t \in [0, t_{f,h} - h]: \\ \tilde{R}_h(t) u_h(t) + \chi_{[s/h]h, t_{f,h}}(t) \tilde{B}_h(t)^* \\ \left\{ \tilde{T}_h(t_{f,h}, t+h)^* G_h \int_{[s/h]h}^{t_{f,h}} \tilde{T}_h(t_{f,h}, \eta+h) \tilde{B}_h(\eta) u_h(\eta) d\eta \right. \\ \left. + \int_{[(t+h)/h]h}^{t_{f,h}} \tilde{T}_h(\eta, t+h)^* \tilde{Q}_h(\eta) \right. \\ \left. \cdot \left[\int_{[s/h]h}^{[\eta/h]h} \tilde{T}_h(\eta, \tau+h) \tilde{B}_h(\tau) u_h(\tau) d\tau \right] d\eta \right\}, \\ \text{For } t \in [t_{f,h} - h, t_{f,h}]: \\ (\tilde{\mathcal{R}}_{h,s} u_h)(t_{f,h} - h), \end{cases}$$

when $s \in [0, t_{f,h})$ and by $(\tilde{\mathcal{R}}_{h,s} u_h)(t) = \tilde{R}_h(t) u_h(t)$, $0 \leq t < t_{f,h}$, when $s \in [t_{f,h}, t_f]$, for $u_h \in \mathcal{U}_h$. Once again, for $j = 0, 1, 2, \dots, k_{f,h} - 1$, $k = j, j + 1, \dots, k_{f,h} - 1$, and $u_{h,j} \in \mathcal{U}_{h,j}$,

$$(2.23) \quad (\tilde{\mathcal{R}}_{h,s} u_h)(t) = (\mathcal{R}_{h,j} u_{h,j})(k)$$

for $s \in [jh, (j+1)h)$ and $t \in [kh, (k+1)h)$, where $u_h \in \mathcal{U}_h$ is given by (2.21) and $\mathcal{R}_{h,j} \in L(\mathcal{U}_{h,j})$ is given by (2.12). The operator $\tilde{\mathcal{R}}_{h,s}$ is self-adjoint and positive definite on \mathcal{U}_h and that if $\tilde{\mathcal{U}}_{h,j}$ denotes the subspace of \mathcal{U}_h obtained from $\mathcal{U}_{h,j}$ via the

natural embedding (i.e., via (2.21)), then, $\tilde{\mathcal{R}}_{h,s}$ is a bijection from $\tilde{\mathcal{U}}_{h,j}$ onto $\tilde{\mathcal{U}}_{h,j}$. It follows therefore from (2.17), (2.19), (2.22) that for $j = 0, 1, 2, \dots, k_{f,h} - 1$

$$(2.24) \quad \mathcal{B}_{h,j}^* \mathcal{R}_{h,j}^{-1} \mathcal{B}_{h,j} \phi = \tilde{\mathcal{B}}_{h,s}^* \tilde{\mathcal{R}}_{h,s}^{-1} \tilde{\mathcal{B}}_{h,s} \phi$$

for each $\phi \in H$ and all $s \in [jh, (j+1)h)$, and that

$$(2.25) \quad \tilde{\mathcal{B}}_{h,s}^* \tilde{\mathcal{R}}_{h,s}^{-1} \tilde{\mathcal{B}}_{h,s} \phi = 0$$

for all $s \in [t_{f,h}, t_f]$.

Setting $\tilde{\Pi}_h(s) = \Pi_h(k)$, $kh \leq s < (k+1)h$, for $s \in [0, t_f]$, from (2.11)–(2.14), and (2.24) we find that

$$(2.26) \quad \begin{aligned} \tilde{\Pi}_h(s) \phi &= \tilde{T}_h(t_{f,h}, s)^* G_h \tilde{T}_h(t_{f,h}, s) \phi \\ &\quad + \int_{[s/h]h}^{t_{f,h}} \tilde{T}_h(t, s)^* \tilde{Q}_h(t) \tilde{T}_h(t, s) \phi dt - \tilde{\mathcal{B}}_{h,s}^* \tilde{\mathcal{R}}_{h,s}^{-1} \tilde{\mathcal{B}}_{h,s} \phi \end{aligned}$$

for each $\phi \in H$. Note that (2.25) implies that $\tilde{\Pi}_h(t) = G_h$ for $t \in [t_{f,h}, t_f]$.

For $\mathcal{B}_s \in L(H, \mathcal{U}_s)$, $\mathcal{R}_s \in L(\mathcal{U}_s)$, and $\mathcal{B}_s^* \in L(\mathcal{U}_s, H)$ given by (2.2)–(2.4), respectively, define $\tilde{\mathcal{B}}_s \in L(H, \mathcal{U})$, $\tilde{\mathcal{R}}_s \in L(\mathcal{U})$, and $\tilde{\mathcal{B}}_s^* \in L(\mathcal{U}, H)$ by

$$(2.27) \quad (\tilde{\mathcal{B}}_s \phi)(t) = \begin{cases} 0, & 0 \leq t < s, \\ (\mathcal{B}_s \phi)(t), & s \leq t \leq t_f; \end{cases}$$

$$(2.28) \quad (\tilde{\mathcal{R}}_s u)(t) = \begin{cases} R(t)u(t), & 0 \leq t < s, \\ (\mathcal{R}_s u)(t), & s \leq t \leq t_f; \end{cases}$$

and

$$(2.29) \quad \tilde{\mathcal{B}}_s^* u = \mathcal{B}_s^* u$$

for $\phi \in H$ and $u \in \mathcal{U}$. Then $\tilde{\mathcal{B}}_s^* = (\tilde{\mathcal{B}}_s)^*$ (i.e., $\tilde{\mathcal{B}}_s^* \in L(\mathcal{U}, H)$ is the Hilbert space adjoint of $\tilde{\mathcal{B}}_s \in L(H, \mathcal{U})$), $\tilde{\mathcal{R}}_s$ is self-adjoint positive definite on \mathcal{U} , and if $\tilde{\mathcal{U}}_s$ denotes the subspace of \mathcal{U} obtained via the natural embedding of \mathcal{U}_s into \mathcal{U} , then $\tilde{\mathcal{R}}_s$ is a bijection from $\tilde{\mathcal{U}}_s$ onto $\tilde{\mathcal{U}}_s$. Consequently, it follows that

$$(2.30) \quad \mathcal{B}_s^* \mathcal{R}_s^{-1} \mathcal{B}_s \phi = \tilde{\mathcal{B}}_s^* \tilde{\mathcal{R}}_s^{-1} \tilde{\mathcal{B}}_s \phi$$

for all $\phi \in H$ and $s \in [0, t_f]$. From (2.5) we obtain that

$$(2.31) \quad \begin{aligned} \Pi(s) \phi &= T(t_f, s)^* G T(t_f, s) \phi + \int_s^{t_f} T(t, s)^* Q(t) T(t, s) \phi dt \\ &\quad - \tilde{\mathcal{B}}_s^* \tilde{\mathcal{R}}_s^{-1} \tilde{\mathcal{B}}_s \phi, \end{aligned}$$

for all $\phi \in H$ and $s \in [0, t_f]$.

Our convergence result for the finite-time horizon problem is given in the following theorem and its corollary.

THEOREM 2.4. *Suppose that the families of operators $\{T, B, Q, R\}$ satisfy conditions (C1)–(C3) and that for all $h > 0$, the families of operators $\{T_h, B_h, Q_h, R_h\}$ satisfy conditions (D1) and (D2). Suppose further that the approximation assumptions (A1)–(A4) are satisfied. Then the discrete-time family of operators $\Pi_h = \{\Pi_h(k) :$*

$0 \leq k \leq k_{f,h}$ given by (2.14) or (2.15) strongly approximates the continuous-time family of operators $\Pi = \{\Pi(t) : 0 \leq t \leq t_f\}$ given by (2.5) or (2.6) on the set $\Delta(1, t_f)$. That is, $\Pi_h \rightarrow \Pi$ on the set $\Delta(1, t_f)$.

Proof. The desired result will follow from Theorem 2.3 if we can argue that $\lim_{h \rightarrow 0+} \tilde{\Pi}_h(t)\phi = \Pi(t)\phi$, uniformly in t , for $t \in (0, t_f]$, for each $\phi \in H$, where $\tilde{\Pi}_h$ and Π are given by (2.26) and (2.31), respectively.

From assumption (A3), we have that the operators $\tilde{\mathcal{R}}_{h,s}^{-1}$ are bounded uniformly in $h > 0$ and $s \in [0, t_f]$. It can be shown that $\tilde{\mathcal{B}}_{h,s}\phi \rightarrow \tilde{\mathcal{B}}_s\phi$, for all $\phi \in H$, $\tilde{\mathcal{R}}_{h,s}P_h u \rightarrow \tilde{\mathcal{R}}_s u$ for $u \in \mathcal{U}$, and $\tilde{\mathcal{B}}_{h,s}^*P_h u \rightarrow \tilde{\mathcal{B}}_s^*u$, for $u \in \mathcal{U}$, uniformly in s for $s \in [0, t_f]$, where $\tilde{\mathcal{B}}_{h,s}$, $\tilde{\mathcal{B}}_s$, $\tilde{\mathcal{R}}_{h,s}$, $\tilde{\mathcal{R}}_s$, $\tilde{\mathcal{B}}_{h,s}^*$, and $\tilde{\mathcal{B}}_s^*$ are given by (2.17), (2.27), (2.22), (2.28), (2.19), and (2.29), respectively (see [RW1]). This, together with the identity

$$\tilde{\mathcal{R}}_{h,s}^{-1}P_h - P_h\tilde{\mathcal{R}}_s^{-1} = \tilde{\mathcal{R}}_{h,s}^{-1}\{\tilde{\mathcal{R}}_{h,s}P_h - P_h\tilde{\mathcal{R}}_s\}\tilde{\mathcal{R}}_s^{-1},$$

yield that $\lim_{h \rightarrow 0+} \tilde{\mathcal{B}}_{h,s}^*\tilde{\mathcal{R}}_{h,s}^{-1}\tilde{\mathcal{B}}_{h,s}\phi = \tilde{\mathcal{B}}_s^*\tilde{\mathcal{R}}_s^{-1}\tilde{\mathcal{B}}_s\phi$, for $\phi \in H$, uniformly in s for $s \in [0, t_f]$. The desired convergence can then be obtained from assumptions (A1), (A2), and (A4) and (2.26) and (2.31). \square

Let $F = \{F(t) : 0 \leq t \leq t_f\}$ and $S = \{S(t, s) : 0 \leq s \leq t \leq t_f\}$ be, respectively, the continuous-time families of optimal closed-loop feedback gain operators and optimal closed-loop state transition operators for the continuous-time LQR problem (P). That is, for $t \in [0, t_f]$

$$F(t) = R(t)^{-1}B(t)^*\Pi(t) \in L(H, \mathcal{U}),$$

and for $0 \leq s \leq t \leq t_f$

$$\begin{aligned} (2.32) \quad S(t, s)\phi &= T(t, s)\phi - \int_s^t T(t, \eta)B(\eta)F(\eta)S(\eta, s)\phi d\eta \\ &= T(t, s)\phi - \int_s^t T(t, \eta)B(\eta)(\mathcal{R}_s^{-1}\mathcal{B}_s^*\phi)(\eta)d\eta, \end{aligned}$$

for $\phi \in H$ (see [G]). Similarly, for the discrete-time problem, let the discrete-time families, $F_h = \{F_h(k) : 0 \leq k \leq k_{f,h} - 1\} \subset L(H, \mathcal{U})$ and $S_h = \{S_h(k, j) : 0 \leq j \leq k \leq k_{f,h}\} \subset L(H)$ be given by

$$F_h(k) = \hat{R}_h(k)^{-1}B_h(k)^*\Pi_h(k+1)A_h(k),$$

where

$$\hat{R}_h(k) = R_h(k) + hB_h(k)^*\Pi_h(k+1)B_h(k),$$

$k = 0, 1, \dots, k_{f,h} - 1$, and

$$\begin{aligned} (2.33) \quad S_h(k, j)\phi &= T_h(k, j)\phi - h \sum_{i=j}^{k-1} T_h(k, i+1)B_h(i)F_h(i)S_h(i, j)\phi \\ &= T_h(k, j)\phi - h \sum_{i=j}^{k-1} T_h(k, i+1)B_h(i)(\mathcal{R}_{h,j}^{-1}\mathcal{B}_{h,j}^*\phi)(i), \end{aligned}$$

$0 \leq j \leq k \leq k_{f,h}$, for $\phi \in H$.

COROLLARY 2.1. *Suppose that the hypotheses of Theorem 2.4 above are satisfied and let $\{\bar{u}, \bar{x}\}$ and $\{\bar{u}_h, \bar{x}_h\}$ be the optimal control/trajectory pairs for the LQR problems (P) and (P_h), respectively, corresponding to the initial data $x(0) = x_h(0) = x_0 \in H$. Then*

- (i) $F_h \rightarrow F$;
- (ii) $S_h \rightarrow S$;
- (iii) $\lim_{h \rightarrow 0^+} \|\bar{u}_h(k_h) - \bar{u}(t)\|_U = 0$, and $\lim_{h \rightarrow 0^+} \|\bar{x}_h(k_h) - \bar{x}(t)\|_H = 0$, for $t \in [0, t_f]$ and for all nets $\{k_h\}_{h>0}$ for which $\lim_{h \rightarrow 0^+} h k_h = t$.
- (iv) $\lim_{h \rightarrow 0^+} \bar{J}_h = \bar{J}$.

Proof. Statements (i) and (iv) (recall (2.7) and (2.16)) are immediate consequences of Theorem 2.4. Statement (iii) follows from statements (i) and (ii) since $\bar{u}(t) = -F(t)\bar{x}(t)$, $\bar{x}(t) = S(t, 0)x_0$, $t \in [0, t_f]$, and $\bar{u}_h(k) = -F_h(k)\bar{x}_h(k)$, $0 \leq k \leq k_{f,h} - 1$, $\bar{x}_h(k) = S_h(k, 0)x_0$, $0 \leq k \leq k_{f,h}$. Thus we need only to verify statement (ii).

We rewrite (2.32) as

$$S(t, s)\phi = T(t, s)\phi - \int_s^t T(t, \eta)B(\eta)(\tilde{\mathcal{R}}_s^{-1}\tilde{\mathcal{B}}_s^*\phi)(\eta)d\eta,$$

and from (2.33) we obtain

$$\tilde{S}_h(t, s)\phi = \tilde{T}_h(t, s)\phi - \int_{[s/h]h}^{[t/h]h} \tilde{T}_h(t, \eta + h)\tilde{B}_h(\eta)(\tilde{\mathcal{R}}_{h,s}^{-1}\tilde{\mathcal{B}}_{h,s}^*\phi)(\eta)d\eta.$$

The result now follows as in the proof of Theorem 2.4. \square

Remark. In actual practice, given the continuous-time LQR problem (P), the net of discrete-time problems $\{(P_h)\}$ is typically obtained by considering zero-order hold (i.e., piecewise constant) control inputs and output sampling. In this case, we would obtain $A_h(k) = T((k+1)h, kh)$, $B_h(k) = h^{-1} \int_{kh}^{(k+1)h} T((k+1)h, s)B(s)ds$, $Q_h(k) = h^{-1} \int_{kh}^{(k+1)h} Q(s)ds$, $R_h(k) = h^{-1} \int_{kh}^{(k+1)h} R(s)ds$, and $G_h = G$. When conditions (C1)–(C3) on the continuous-time families T, B, Q , and R are satisfied, it is immediately clear that the discrete-time families T_h, B_h, Q_h and R_h , and the operator G_h satisfy conditions (D1) and (D2) and the approximation conditions (A1)–(A4). More generally, other discretizations are also admissible. For example, in the time-invariant case, the semigroup $\{T(t) : t \geq 0\}$ could be discretely approximated using A -stable Padé approximants to the exponential (see [HK]). In particular, if $T(t) = \exp(tA)$, $t \geq 0$, then we might set $T_h(k) = (I - hA)^{-k}$ (implicit Euler) or $T_h(k) = (I - hA/2)^{-k}(I + hA/2)^k$ (Crank–Nicolson). The stability and consistency of these discretizations (i.e., assumption (A2)) can be verified using the theory and techniques developed in [HK].

3. The infinite-time horizon problem. In the LQR problem over an infinite-time interval, the state equations (2.1) and (2.9) governing the dynamics of the continuous-time and discrete-time control systems, respectively, remain the same. The continuous- and discrete-time operator families $\{T, B, Q, R\}$, and $\{T_h, B_h, Q_h, R_h\}$ are assumed to be defined on the infinite-time intervals $[t_0, +\infty) \subset \mathbb{R}$ and $[k_0, +\infty) \subset \mathbb{Z}$, respectively. The cost functionals are taken to be

$$\begin{aligned} (3.1) \quad J_\infty(u; t_0, x(t_0)) &= \int_{t_0}^\infty \{ \langle Q(t)x(t), x(t) \rangle_H + \langle R(t)u(t), u(t) \rangle_U \} dt \\ &= \lim_{t_f \rightarrow \infty} J(u; t_0, x(t_0), 0) \end{aligned}$$

and

$$\begin{aligned}
 J_{h,\infty}(u_h; k_0, x_h(k_0)) &= h \sum_{k=k_0}^{\infty} \{ \langle Q_h(k)x_h(k), x_h(k) \rangle_H + \langle R_h(k)u_h(k), u_h(k) \rangle_U \} \\
 (3.2) \qquad \qquad \qquad &= \lim_{k_f \rightarrow \infty} J_h(u_h; k_0, x_h(k_0), 0)
 \end{aligned}$$

Under the usual stabilizability and detectability assumptions on the continuous-time and the discrete-time control systems, the existence and the uniqueness of the optimal controls \bar{u}, \bar{u}_h minimizing (3.1) and (3.2), respectively, can be guaranteed. Moreover, these optimal controls can be written in a closed-loop state-feedback form (see Theorem 3.1 below). We are again interested in investigating the convergence of the optimal controls and the optimal feedback laws for the sampled systems as the length of the sampling interval tends toward zero. Once again, for simplicity, we assume henceforth, without loss of generality, that $t_0 = k_0 = 0$.

Our fundamental convergence result in this case can be summarized as follows. Assume that the conditions (A1)–(A4) are satisfied on every finite-time interval $[t_0, t_f]$. Suppose further that the stabilizability and the detectability of the continuous-time system are uniformly preserved by the sampled-time systems (see Definitions 3.3(iii) and 3.4(iii) below). Then the optimal controls \bar{u}_h and the optimal state-feedback laws F_h for the sampled-time systems converge to the optimal control \bar{u} and optimal feedback law F for the continuous-time system, respectively, as the length h of the sampling interval tends toward zero.

This result is a direct consequence of the main result of this section, Theorem 3.4, which is concerned with the convergence of the associated Riccati operators. Theorem 3.4 is obtained by first establishing a convergence result that requires the somewhat difficult to verify condition of uniform exponential stability of the optimal discrete-time closed-loop systems, uniformly in the sampling rate (Theorem 3.2 below). Theorem 3.3, which is of some interest in its own right, is primarily used in the verification of the conditions of Theorem 3.4. To make our presentation complete and self contained, the well-known existence and uniqueness result for the closed-loop linear state-feedback solution to continuous- and discrete-time LQR problems on the infinite interval are stated in Theorem 3.1.

DEFINITION 3.1. (Cost functional stabilizability)

(i) The continuous-time system associated with the operator pair $\{T, B\}$ is said to be cost-functional stabilizable with respect to the performance index J_∞ given by (3.1), if for each $\phi \in H$, there exists a constant $M(\phi)$ such that for any $s \geq 0$, there exists a control input $u_s \in L_2(s, \infty; U)$ with $J_\infty(u_s; s, \phi) \leq M(\phi)$.

(ii) The sampled-time system associated with the operator pair $\{T_h, B_h\}$ is said to be cost-functional stabilizable with respect to the discrete performance index $J_{h,\infty}$ given by (3.2), if for each $\phi \in H$, there exists a constant $M_h(\phi)$ such that for any $j \geq 0$, there exists a control input sequence $u_{h,j} \in l_2(j, \infty; U)$ with $J_{h,\infty}(u_{h,j}; j, \phi) \leq M_h(\phi)$.

(iii) The sampled systems are said to be uniformly cost-functional stabilizable for all $0 < h \leq h_0$, if for each $\phi \in H$, the constants $M_h(\phi)$ defined in (ii) are independent of the length of the sampling interval h , for all $h \leq h_0$ for some $h_0 > 0$.

For any given final time t_f and final index $k_{f,h}$, let $\Pi_{t_f}(\cdot; G)$ and $\Pi_{k_{f,h}}(\cdot; G_h)$ denote the Riccati operators given by (2.5) and (2.14) corresponding to the final state penalty operators G and G_h , respectively. In the case where $G = G_h = 0$, using (2.7) and (2.16), it is easy to verify that (see for example, [DI]) for each given $t \geq 0$ and $k \geq 0$, the functions $t_f \mapsto \Pi_{t_f}(t; 0)$ and $k_{f,h} \mapsto \Pi_{k_{f,h}}(k; 0)$ are nondecreasing,

selfadjoint, nonnegative operator-valued functions. If cost functional stabilizability of the continuous- and discrete-time control systems is assumed then Π_{t_f} and $\Pi_{h,k_f,h}$ are bounded above. Indeed, we have

$$\langle \Pi_{t_f}(t;0)\phi, \phi \rangle_H \leq M(\phi), \quad \phi \in H,$$

and

$$\langle \Pi_{h,k_f,h}(k;0)\phi, \phi \rangle_H \leq M_h(\phi), \quad \phi \in H,$$

for all t_f and $k_{f,h}$. Thus, strong limits of $\Pi_{t_f}(t;0)$ and $\Pi_{h,k_f,h}(k;0)$ exist for each $t \geq 0$ and $k \geq 0$ as t_f and $k_{f,h}$ tend to infinity. We denote these strong limiting operator-valued functions by $\Pi_\infty(\cdot;0)$ and $\Pi_{h,\infty}(\cdot;0)$, respectively. The existence and uniqueness of the solutions to the continuous- and discrete-time optimal control problems is given in the following well-known theorem; see, for example, [BW], [G], [GR], [LCB], [HH], and [Z].

THEOREM 3.1. *Assume that the continuous-time system and the sampled time systems for all h sufficiently small are cost-functional stabilizable. Then for any $s \geq 0$ and $j \geq 0$, and initial states $x(s) = \phi$ and $x_{j,h} = \phi$, there exist unique optimal controls \bar{u} and \bar{u}_h , which minimize the cost functionals $J_\infty(\cdot; s, x(s); 0)$ over $L_2(s, \infty; U)$ and $J_{h,\infty}(\cdot; j, x_h(j); 0)$ over $l_2(j, \infty; U)$, respectively. The optimal controls can be written in linear state feedback form as*

$$\bar{u}(t) = -R(t)^{-1}B(t)^*\Pi_\infty(t;0)\bar{x}(t) = -F(t)\bar{x}(t),$$

and

$$\bar{u}_h(k) = -\hat{R}_h(k)^{-1}B_h(k)^*\Pi_{h,\infty}(k+1;0)A_h(k)\bar{x}_h(k) = -F_h(k)\bar{x}_h(k),$$

where \bar{x} and \bar{x}_h are the corresponding optimal trajectories and $\hat{R}_h(k) = R_h(k) + hB_h(k)^*\Pi_{h,\infty}(k+1;0)B_h(k)$. The operator-valued function $\Pi_\infty(\cdot;0)$ is bounded on the interval $[0, \infty)$ and satisfies the Riccati integral equation

$$(3.3) \quad \begin{aligned} \Pi_\infty(s;0)\phi &= T(t,s)^*\Pi_\infty(t;0)T(t,s)\phi \\ &+ \int_s^t T(\tau,s)^*[Q(\tau) - \Pi_\infty(\tau;0)(BR^{-1}B^*)(\tau)\Pi_\infty(\tau;0)]T(\tau,s)\phi d\tau, \end{aligned}$$

for all $\phi \in H$ and $(t,s) \in \Delta(2,\infty)$. Similarly, the operator-valued sequence $\Pi_{h,\infty}(\cdot;0)$ is bounded for $0 \leq k < \infty$ and satisfies the Riccati difference equation

$$(3.4) \quad \begin{aligned} \Pi_{h,\infty}(k;0) &= A_h(k)^*\Pi_{h,\infty}(k+1;0)A_h(k) + hQ_h(k) \\ &- hA_h(k)^*\Pi_{h,\infty}(k+1;0)B_h(k)\hat{R}_h(k)^{-1}B_h(k)^*\Pi_{h,\infty}(k+1;0)A_h(k). \end{aligned}$$

If the sampled-time systems are uniformly cost-functional stabilizable for $0 < h \leq h_0$, then the operator-valued sequences $\Pi_{h,\infty}(\cdot;0)$ are uniformly bounded for all sampling periods h with $0 < h \leq h_0$.

If it is assumed that the approximation conditions (A1)–(A4) hold, then from Theorem 3.1, it is not difficult to see that on a given finite-time interval $[0, t_f]$, the uniform convergence of the optimal controls \bar{u}_h , the optimal trajectories \bar{x}_h , and the optimal feedback gains F_h for the sampled-time control problems would follow directly from the uniform convergence of $\Pi_{h,\infty}(\cdot;0)$. Our investigation is, therefore, focused on

the convergence of $\Pi_{h,\infty}(\cdot; 0)$ to $\Pi_\infty(\cdot; 0)$ as h tends toward zero. Using the notation introduced in the previous section, we note that for each $t \geq 0$, an obvious sufficient condition for the convergence of $\tilde{\Pi}_{h,\infty}(t; 0)$ to $\Pi_\infty(t; 0)$ is the convergence of $\Pi_{t_f}(t; G)$ to $\Pi_\infty(t; 0)$ and the uniform convergence in h of $\tilde{\Pi}_{h,k_{f,h}}(t; G_h)$ to $\tilde{\Pi}_{h,\infty}(t; 0)$ (with $k_{f,h} = [t_f/h]$) as t_f tends to infinity for some $G \geq 0$ and corresponding $G_h \geq 0$. Indeed, from the triangle inequality, for $\phi \in H$, we have

$$\begin{aligned} \|\tilde{\Pi}_{h,\infty}(t; 0)\phi - \Pi_\infty(t; 0)\phi\|_H &\leq \|\tilde{\Pi}_{h,\infty}(t; 0)\phi - \tilde{\Pi}_{h,k_{f,h}}(t; G_h)\phi\|_H \\ &+ \|\tilde{\Pi}_{h,k_{f,h}}(t; G_h)\phi - \Pi_{t_f}(t; G)\phi\|_H + \|\Pi_{t_f}(t; G)\phi - \Pi_\infty(t; 0)\phi\|_H. \end{aligned}$$

Then for an arbitrary $\epsilon > 0$, a sufficiently large t_f can be chosen such that the first and the last terms on the right-hand side of the above inequality are smaller than $\epsilon/3$ for all h . By applying the theory of the previous section on the interval $[0, t_f]$, there exists $h_0 > 0$ small enough such that for all $0 < h \leq h_0$, the second term on the right hand side of the above inequality is bounded by $\epsilon/3$. Thus, the desired convergence immediately follows.

If the trajectories of the discrete- and continuous-time systems are asymptotically stable, then as t_f tends to infinity, the cost functionals J_∞ and $J_{h,\infty}$ are also limits of the cost functionals J, J_h for the finite-time interval problems on $[0, t_f]$ with final state penalties G and G_h different from zero. In particular, if the optimal trajectory of the infinite-horizon problem is asymptotically stable, the convergence rates of $\bar{J}_h(\bar{u}_h; k, \phi, G_h) = \langle \Pi_{h,k_{f,h}}(k; G_h)\phi, \phi \rangle_H$ with $G_h \geq M_h(\phi)$ and $\bar{J}(\bar{u}; t, \phi, G) = \langle \Pi_{t_f}(t; G)\phi, \phi \rangle_H$ with $G \geq M(\phi)$ can be estimated by the decay rate of the optimal trajectory \bar{x} for the infinite-horizon problem. Toward this end, let $S = \{S(t, s) : 0 \leq s \leq t < \infty\}$ be the continuous-time evolution system given by

$$(3.5) \quad S(t, s)\phi = T(t, s)\phi - \int_s^t T(t, \tau)B(\tau)F(\tau)S(\tau, s)\phi ds, \quad \text{for } \phi \in H.$$

The evolution system S is also referred to as the perturbation of T by $-BF$. It is not difficult to verify that $S(t, 0)\phi$ corresponds to the optimal trajectory for the continuous-time infinite-horizon problem with initial state $\phi \in H$. Similarly, let the discrete-time evolution system $S_h = \{S_h(i, j) : 0 \leq j \leq i < \infty\}$ be defined as

$$(3.6) \quad S_h(i, j) = \begin{cases} I & i = j, \\ \prod_{k=j}^{i-1} \{A_h(k) - hB_h(k)\hat{R}_h(k)^{-1}B_h(k)^* \Pi_{h,\infty}(k+1; 0)A_h(k)\}, & i > j. \end{cases}$$

Thus, $S_h(k, 0)\phi$ is the optimal trajectory for the discrete-time infinite-horizon problem with initial state $\phi \in H$.

DEFINITION 3.2. (Exponential stability of the optimal feedback systems)

(i) The optimal continuous-time feedback system (3.5) is said to be exponentially stable, if there exist constants M and $\alpha > 0$ such that for all $0 \leq s \leq t < \infty$, $\|S(t, s)\|_{L(H)} \leq M \exp\{-\alpha(t-s)\}$.

(ii) The discrete-time optimal feedback system (3.6) is said to be exponentially stable, if there exist constants M_h and $\alpha_h > 0$ such that, for all $0 \leq j \leq i < \infty$, $\|S_h(i, j)\|_{L(H)} \leq M_h \exp\{-\alpha_h(i-j)h\}$.

(iii) The sampled-time optimal feedback systems are said to be uniformly exponentially stable for all $0 < h \leq h_0$, if the constants M_h and $\alpha_h > 0$ in (ii) above are independent of h for $0 < h \leq h_0$.

Our first convergence result is given in Theorem 3.2 below. To establish it we require the following two lemmas. The first of these lemmas is an important property of the solutions of the Riccati equations on the infinite-time interval when the optimal feedback systems are exponentially stable. The proof can be found in [BW], [DI], [G] for the continuous time problem, and in [GR, Thm. 2.9] for the discrete-time problem.

LEMMA 3.1. *Assume that the continuous-time control system and the sampled-time control system with sampling period h are cost-functional stabilizable. If the corresponding optimal feedback systems are exponentially stable, then $\Pi_\infty(\cdot; 0)$, and $\Pi_{h,\infty}(\cdot; 0)$ are the unique bounded solutions of the corresponding Riccati equations (3.3) and (3.4) on the infinite-time interval. Furthermore, if G and G_h are chosen such that $G \geq \Pi_\infty(t; 0)$ and $G_h \geq \Pi_{h,\infty}(k; 0)$ for all t and k , then the solutions of the Riccati equations on the finite-time interval, $\Pi_{t_f}(t; G)$ and $\Pi_{h,k_{f,h}}(k; G_h)$, satisfy*

$$\langle \Pi_{t_f}(t; G)\phi - \Pi_\infty(t; 0)\phi, \phi \rangle_H \leq \langle GS(t_f, t)\phi, S(t_f, t)\phi \rangle_H,$$

and

$$\langle \Pi_{h,k_{f,h}}(k; G_h)\phi - \Pi_{h,\infty}(k; G_h)\phi, \phi \rangle_H \leq \langle G_h S_h(k_{f,h}, k)\phi, S_h(k_{f,h}, k)\phi \rangle_H,$$

respectively, for all $t \leq t_f$, $k \leq k_{f,h}$, and $\phi \in H$.

LEMMA 3.2. *Assume that the sampled systems are uniformly cost-functional stabilizable with the optimal feedback systems uniformly exponentially stable for $0 < h \leq h_0$. Then, the operators G and G_h can be chosen as described in Lemma 3.1 with $G_h \leq C \cdot I$ for some constant C independent of h . As t_f tends to infinity, $\Pi_{t_f}(\cdot; G)$ converges to $\Pi_\infty(\cdot; 0)$ uniformly on any bounded subinterval $[a, b]$ of $[0, \infty)$ and the convergence of $\Pi_{h,k_{f,h}}(\cdot; G_h)$ with $k_{f,h} = [t_f/h]$ to $\Pi_{h,\infty}(\cdot; 0)$ is uniform in h for all $0 < h \leq h_0$ on any bounded subinterval $[a, b]$ of $[0, \infty)$ in the uniform operator norm.*

Proof. We prove only the discrete-time assertion. The continuous-time case is completely analogous, if not simpler. The assumption of uniform cost-functional stabilizability implies that the operators G_h can be chosen as stated in the theorem. Then let M and α be the constants in Definition 3.2(iii). For a given $\epsilon > 0$ and $t \in [a, b]$, we can take t_f large enough such that $CM^2 \exp\{-2\alpha(t_f - t - h_0)\} \leq \epsilon$. Let $k_h = [t/h]$, then $(k_{f,h} - k_h)h \geq t_f - t - h_0$ for all $0 < h \leq h_0$. Since $\Pi_{h,k_{f,h}}(k_h; G_h) \geq \Pi_{h,\infty}(k_h; 0)$, using the previous lemma we find that

$$\begin{aligned} \|\tilde{\Pi}_{h,k_{f,h}}(t; G_h) - \tilde{\Pi}_{h,\infty}(t; 0)\|_{L(H)} &= \|\Pi_{h,k_{f,h}}(k_h; G_h) - \Pi_{h,\infty}(k_h; 0)\|_{L(H)} \\ &= \sup_{\|\phi\|_H \leq 1} \langle (\Pi_{h,k_{f,h}}(k_h; G_h) - \Pi_{h,\infty}(k_h; 0))\phi, \phi \rangle_H \\ &\leq \sup_{\|\phi\|_H \leq 1} \langle G_h S_h(k_{f,h}, k_h)\phi, S_h(k_{f,h}, k_h)\phi \rangle_H \\ &\leq CM^2 e^{-2\alpha(k_{f,h} - k_h)h} \leq \epsilon. \quad \square \end{aligned}$$

THEOREM 3.2. *Assume that conditions (A1)–(A4) for the operator families $\{T_h, B_h, Q_h, R_h\}$ hold on any finite subinterval of $[0, \infty)$. Assume further that the continuous-time system and the sampled-time systems with $0 < h \leq h_0$ are uniformly cost-functional stabilizable, and that the optimal closed-loop evolution systems are uniformly exponentially stable. Then, the Riccati operators $\Pi_{h,\infty}(t; 0)$ converge strongly to $\Pi_\infty(t; 0)$ and the convergence is uniform on any bounded subinterval of $[0, \infty)$.*

Proof. Let $\phi \in H$ and let $[a, b]$ be a bounded subinterval of $[0, \infty)$. We choose an operator G such that $G \geq \Pi_\infty(t; 0)$ and $G \geq \Pi_{h,\infty}(k; 0)$ for all $t \in [0, \infty) \subset R$,

$k \in [0, \infty) \subset Z$ and $0 < h \leq h_0$. By Lemma 3.2, t_f can be taken large enough such that for all $k_{f,h} = [t_f/h]$, we have

$$\|\Pi_{t_f}(t; G)\phi - \Pi_\infty(t; 0)\phi\|_H \leq \frac{\epsilon}{3} \quad \text{and} \quad \|\tilde{\Pi}_{h,k_{f,h}}(t; G)\phi - \tilde{\Pi}_{h,\infty}(t; 0)\phi\|_H \leq \frac{\epsilon}{3},$$

for all $t \in [a, b]$ and all $0 < h \leq h_0$. By Theorem 2.4 of §2, we can find h small enough such that

$$\|\tilde{\Pi}_{h,k_{f,h}}(t; G)\phi - \Pi_{t_f}(t; G)\phi\|_H \leq \frac{\epsilon}{3},$$

for all $t \in [a, b]$. Therefore, we have

$$\begin{aligned} \|\tilde{\Pi}_{h,\infty}(t; 0)\phi - \Pi_\infty(t; 0)\phi\|_H &\leq \|\tilde{\Pi}_{h,\infty}(t; 0)\phi - \tilde{\Pi}_{h,k_{f,h}}(t; G)\phi\|_H \\ &+ \|\tilde{\Pi}_{h,k_{f,h}}(t; G)\phi - \Pi_{t_f}(t; G)\phi\|_H + \|\Pi_{t_f}(t; G)\phi - \Pi_\infty(t; 0)\phi\|_H \leq \epsilon, \end{aligned}$$

for all $t \in [a, b]$. \square

The discussion to follow is concerned with conditions that guarantee the uniform exponential stability of the optimal feedback systems. A useful characterization of exponentially stable evolution systems is given in a result due to Datko in the continuous-time case (see [D]) and Zabczyk in the discrete-time case (see [Z]). We state it here in both its continuous- and discrete-time forms as a lemma.

LEMMA 3.3. (i) *Let T be a strongly continuous evolution system. If there exists constants C_1, C_2 , and $\omega > 0$ such that*

$$\|T(t, s)\|_{L(H)} \leq C_1 e^{\omega(t-s)} \quad \text{and} \quad \int_s^\infty \|T(t, s)\phi\|_H^2 dt \leq C_2 \|\phi\|_H^2,$$

for all $\phi \in H$ and $0 \leq s \leq t < \infty$, then we can find constants M and $\alpha > 0$, depending only on C_1, C_2 , and ω , such that $\|T(t, s)\|_{L(H)} \leq M \exp\{-\alpha(t-s)\}$, for all $0 \leq s \leq t < \infty$.

(ii) *Let T_h be the discrete-time evolution system defined by*

$$T_h(i, j) = \begin{cases} I, & i = j, \\ \prod_{k=j}^{i-1} A_h(k), & i > j. \end{cases}$$

If there exist constants $C_{1,h}, \omega_h$ and $C_{2,h}$ such that

$$\|T_h(i, j)\|_{L(H)} \leq C_{1,h} e^{\omega_h(i-j)h} \quad \text{and} \quad h \sum_{i=k}^\infty \|T_h(i, k)\phi\|_H^2 \leq C_{2,h} \|\phi\|_H^2,$$

for all $0 \leq k < \infty$ and $\phi \in H$, then we can find constants M_h and $\alpha_h > 0$, depending only on $C_{1,h}, \omega_h$ and $C_{2,h}$, such that for all $0 \leq j \leq i < \infty$

$$\|T_h(i, j)\|_{L(H)} \leq M_h e^{-\alpha_h(i-j)h}.$$

If the operators $Q(t)$ and $Q_h(k)$ are uniformly strictly coercive (i.e., there exists a constant $q > 0$, such that $Q(t) \geq qI$ and $Q_h(k) \geq qI$, for $t \geq 0$ and $k \geq 0$), it is easily shown that (see [RW1]) S and S_h are uniformly exponentially stable. More generally, the cost functional being bounded implies the stability of the feedback system when

the system is detectable. We define the notion of detectability and then establish this result in Theorem 3.3 to follow.

DEFINITION 3.3. (Detectability)

(i) A continuous-time control system is said to be detectable with respect to the cost functional (3.1) if there exists a bounded operator-valued function $V(\cdot) : [0, \infty) \mapsto L(H)$ such that the evolution system T_V , corresponding to the perturbation of T by $VQ^{1/2}(\cdot)$, is exponentially stable.

(ii) A sampled-time control system is said to be detectable with respect to the cost functional (3.2), if there exists a bounded sequence of operators $\{V_h(k)\}_{k=0}^\infty \subset L(H)$ such that the discrete-time evolution system $T_{V,h}$ given by

$$T_{V,h}(i, j) = \begin{cases} I, & i = j, \\ \prod_{k=j}^{i-1} (A_h(k) + hV_h(k)Q_h(k)^{1/2}), & i > j, \end{cases}$$

is exponentially stable.

(iii) The sampled-time systems are said to be uniformly detectable for $0 < h \leq h_0$, if there exist constants C_1, C_2 , and $\alpha > 0$, independent of h such that the operator-valued sequences $\{V_h(k)\}_{k=0}^\infty$ in (3.3) satisfy $\|V_h(k)\|_{L(H)} \leq C_1$ and

$$\|T_{V,h}(i, j)\|_{L(H)} \leq C_2 e^{-\alpha(i-j)h}, \quad 0 \leq j \leq i < \infty,$$

for all sampling rates $0 < h \leq h_0$.

In what follows we shall also require the following assumption.

(B) The continuous-time evolution system T and the discrete-time evolution system T_h are uniformly exponentially bounded on $\Delta(2, \infty)$. That is, there exist constants M and ω such that

$$\|T(t, s)\|_{L(H)} \leq M e^{\omega(t-s)}, \quad \|T_h(i, j)\|_{L(H)} \leq M e^{\omega(i-j)h},$$

for $0 \leq s \leq t < \infty$ and $0 \leq j \leq i < \infty$. The operator families B, Q , and R and the piecewise constant operator families \tilde{B}_h, \tilde{Q}_h , and \tilde{R}_h are uniformly bounded in norm by a given constant C on the entire interval $[0, \infty)$ for all sampling rates $h > 0$. Furthermore, there exists a constant $r > 0$ such that $R(t) \geq rI$ and $\tilde{R}_h(t) \geq rI$ for all $t \geq 0$, and $h > 0$.

THEOREM 3.3. Consider a detectable continuous-time control system and a detectable sampled-time system that are both cost-functional stabilizable. Assume that the evolution systems T, T_h are exponentially bounded, and the operator families $\{B, Q, R\}$ and $\{B_h, Q_h, R_h\}$ are bounded in norm on the infinite-time interval. Then, the optimal feedback systems for both systems are exponentially stable. Furthermore, suppose that constants $C, \omega, r > 0$, and $\alpha > 0$ can be found such that the following conditions are satisfied.

(i) The operator families $\{B, Q, R, \Pi_\infty(\cdot; 0), V\}$ and $\{B_h, Q_h, R_h, \Pi_{h,\infty}(\cdot; 0), V_h\}$ are bounded in norm by C ;

(ii) For all $t \geq 0, k \geq 0$, $R(t) \geq rI$, and $R_h \geq rI$;

(iii) The evolution systems T, T_h, T_V , and $T_{V,h}$ satisfy

$$\|T(t, s)\|_{L(H)} \leq C e^{\omega(t-s)}, \quad \|T_V(t, s)\|_{L(H)} \leq C e^{-\alpha(t-s)},$$

and

$$\|T_h(i, j)\|_{L(H)} \leq C e^{\omega(i-j)h}, \quad \|T_V(i, j)\|_{L(H)} \leq C e^{-\alpha(i-j)h}.$$

Then there exist constants M and $\beta > 0$ depending only on the constants C, τ, α , and ω such that

$$\|S(t, s)\|_{L(H)} \leq Me^{-\beta(t-s)}, \quad \|S_h(i, j)\|_{L(H)} \leq Me^{-\beta(i-j)h}.$$

Moreover, under assumption (B), if the sampled systems are uniformly detectable and uniformly cost-functional stabilizable for $0 < h \leq h_0$, then the optimal closed-loop systems are uniformly exponentially stable for $0 < h \leq h_0$.

Proof. In the case of continuous-time system, a proof is given by Da Prato and Ichikawa in [DI]. The dependence of the exponential bound for the optimal closed-loop system on the constants indicated above is proved in [W]. The arguments for the discrete-time case are very similar to those used in the continuous-time case. Indeed, let S_h correspond to the perturbation of $T_{V,h}$ by $\Delta_h = \{\Delta_h(k) = -B_h(k)F_h(k) + V_h(k)Q_h(k)^{1/2}\}$ in the sense that

$$\tilde{S}_h(t, s)\phi = \tilde{T}_{V,h}(t, s)\phi + \int_{[s/h]}^{[t/h]} \tilde{T}_{V,h}(t, \tau) \tilde{\Delta}_h(\tau) \tilde{S}_h(\tau, s)\phi d\tau.$$

Let us define

$$f_h(k, i) = -R_h(k)^{1/2}F_h(k)\phi \quad \text{and} \quad g_h(k, i) = Q_h(k)^{1/2}S_h(k, i)\phi,$$

for $k \geq i \geq 0$. Then cost-functional stabilizability implies that

$$\|f_h(\cdot, i)\|_{l_2(i, \infty; U)} \leq C\|\phi\|_H, \quad \|g_h(\cdot, i)\|_{l_2(i, \infty; H)} \leq C\|\phi\|_H.$$

The evolution system $T_{V,h}$ is bounded; $\|T_{V,h}(i, j)\|_{L(H)} \leq C \exp\{-\alpha(i-j)h\}$. Thus we obtain

$$\begin{aligned} \|\tilde{S}_h(t, s)\phi\|_H &\leq Ce^{-\alpha(t-s)} + \int_{[s/h]}^{[t/h]} Ce^{-\alpha(t-\tau)} (\|\tilde{B}_h(\tau)\tilde{R}_h(\tau)^{-1/2}\|_{L(U, H)} \|\tilde{f}_h(\tau, s)\|_U \\ &\quad + \|\tilde{V}_h(\tau)\|_{L(H)} \|\tilde{g}_h(\tau, s)\|_H) d\tau, \end{aligned}$$

and by Young's inequality (see [A, Theorem 4.30, p. 90]), we have

$$\int_s^\infty \|\tilde{S}_h(t, s)\phi\|_H^2 dt \leq K\|\phi\|_H^2, \quad \phi \in H,$$

for some constant K . Applying Lemma 3.3, we obtain the exponential stability of S_h . The dependence of the exponential bound for S_h on the indicated constants of course follows from the dependence of the constant K on the indicated constants as prescribed in the lemma. In this way it is easy to see how under assumption (B), uniform detectability and cost function stabilizability will imply the uniform exponential stability of the closed-loop systems. \square

Dual to detectability is the notion of stabilizability.

DEFINITION 3.4. (Stabilizability)

(i) A continuous-time system is said to be stabilizable if there exists a bounded operator-valued function $K(\cdot) : [t_0, \infty) \mapsto L(H, U)$ such that the evolution system T_K corresponding to the perturbation of T by BK is exponentially stable. (ii) A sampled system is said to be stabilizable if there exists a bounded sequence of operators $\{K_h(k)\}_{k=0}^\infty \subset L(H, U)$ such that the discrete evolution operator $T_{K,h}$ given by

$$T_{K,h}(i, j) = \begin{cases} I, & i = j, \\ \prod_{k=j}^{i-1} (A_h(k) + hB_h(k)K_h(k)), & i > j, \end{cases}$$

is exponentially stable. (iii) The sampled-time systems for are said to be uniformly stabilizable for $0 < h \leq h_0$ if there exist constants $C_1, C_2, \alpha > 0$ independent of the sampling period h , such that K_h and $T_{K,h}$ satisfy

$$\|K_h(k)\|_{L(H,U)} \leq C_1, \quad \|T_{K,h}(i,j)\|_{L(H)} \leq C_2 e^{-\alpha(i-j)h},$$

for all $0 \leq k < \infty, 0 \leq j \leq i < \infty$.

Using Theorem 3.3, it is easy to verify that cost-functional stabilizability and detectability imply stabilizability (take $K = F, K_h = F_h$, for example). Conversely, stabilizability clearly implies cost-functional stabilizability. Therefore, under the uniform detectability assumption, cost-functional stabilizability and stabilizability are equivalent. In general, uniform stabilizability and uniform detectability are required for the convergence of $\Pi_{h,\infty}$ to Π_∞ as h tends toward zero. Thus we have our second, and more useful, convergence result.

THEOREM 3.4. *Let assumption (B) hold. Suppose further that Conditions (A1)–(A4) hold on any bounded subinterval of $[0, \infty)$. If the continuous-time system and the sampled-time systems are uniformly stabilizable and uniformly detectable, then the unique solution $\Pi_{h,\infty}$ of the infinite-horizon Riccati difference equation (3.4) converges to the solution Π_∞ of the infinite-horizon Riccati integral equation (3.3) as h tends toward zero. The convergence is uniform in time on any bounded subinterval of $[0, \infty)$.*

Proof. By Theorem 3.3, uniform stabilizability and uniform detectability imply exponential stability of the optimal feedback systems (i.e., Definition 3.2), uniformly over all sampled systems with $0 < h \leq h_0$. Therefore, by Theorem 3.2, we obtain the desired convergence. \square

4. Finite rank and uniform stabilizability and detectability. Most control systems of interest in engineering practice are stabilizable and detectable. In fact, in modeling many control systems of practical interest, a realistic description of the physical system frequently necessitates stabilizability and detectability of the system model (see, for example, [BKS], [BKSW]). Investigation of stabilizability and detectability of particular classes of evolution systems has generated several interesting mathematical problems (see, for example, [C], [L]). However, in the context of approximation, we usually assume that the original control system is stabilizable and detectable. As we have seen in the previous section, the important issue here is whether a given time discretization algorithm is capable of preserving these properties uniformly in the sampling rate, and therefore provide convergent discrete-time approximations for the optimal feedback operators. In this section, we attempt to address this issue for some particular discretization algorithms and derive sufficient conditions for uniform stabilizability and detectability. These conditions take the form of what we call finite rank stabilizability and detectability.

Assume that the control system defined in (2.1) is stabilizable and detectable with respect to the cost functional (3.1). Thus, there exist bounded operator-valued functions $K(\cdot) : [t_0, \infty) \mapsto L(H, U)$ and $V(\cdot) : [t_0, \infty) \mapsto L(H)$ such that the evolution systems T_K, T_V , corresponding to the perturbations of T by BK and $VQ^{1/2}$, respectively, are exponentially stable. That is, there exist constants $M, \alpha > 0$ such that $\|T_K(t, s)\|_{L(H)} \leq M \exp\{-\alpha(t-s)\}$ and $\|T_V(t, s)\|_{L(H)} \leq M \exp\{-\alpha(t-s)\}$, for all $0 \leq s \leq t < \infty$. By definition, the evolution operators T_K and T_V satisfy

$$(4.1) \quad T_K(t, s)\phi = T(t, s)\phi + \int_s^t T(t, \eta)B(\eta)K(\eta)T_K(\eta, s)\phi d\eta,$$

$$(4.2) \quad T_V(t, s)\phi = T(t, s)\phi + \int_s^t T(t, \eta)V(\eta)Q^{1/2}(\eta)T_V(\eta, s)\phi d\eta,$$

for all $\phi \in H$ and for all $0 \leq s \leq t < \infty$. Consider the zero-order hold discretization described in §2. For each $k \geq 0$, the operators $A_h(k), B_h(k)$ are defined by

$$(4.3) \quad A_h(k) = T((k+1)h, kh),$$

$$(4.4) \quad B_h(k) = \frac{1}{h} \int_{kh}^{(k+1)h} T((k+1)h, \eta)B(\eta)d\eta,$$

with the discrete evolution systems $T_{K,h}, T_{V,h}$ then given by

$$(4.5) \quad T_{K,h}(i, j) = \begin{cases} I, & i = j, \\ \prod_{k=j}^{i-1} \{A_h(k) + hB_h(k)K(kh)\}, & i > j, \end{cases}$$

$$(4.6) \quad T_{V,h}(i, j) = \begin{cases} I, & i = j, \\ \prod_{k=j}^{i-1} \{A_h(k) + hV(kh)Q_h(k)^{1/2}\}, & i > j. \end{cases}$$

If the discrete-time evolution systems $T_{K,h}, T_{V,h}$ are uniformly exponentially stable for all $0 < h \leq h_0$ for some $h_0 > 0$, then these sampled-time systems are uniformly stabilizable and uniformly detectable. Using (4.1) and (4.2), the evolution systems T_K and T_V satisfy

$$T_K(ih, jh) = \prod_{k=j}^{i-1} \left[T((k+1)h, kh) + \int_{kh}^{(k+1)h} T((k+1)h, \eta)B(\eta)K(\eta)T_K(\eta, kh)d\eta \right]$$

and

$$T_V(ih, jh) = \prod_{k=j}^{i-1} \left[T((k+1)h, kh) + \int_{kh}^{(k+1)h} T((k+1)h, \eta)V(\eta)Q(\eta)^{1/2}T_V(\eta, kh)d\eta \right],$$

for $0 \leq j \leq i < \infty$. Therefore, $T_{K,h}$ and $T_{V,h}$, given in (4.5) and (4.6) above, can be considered as perturbations of T_K and T_V , respectively. In fact, we have

$$(4.7) \quad T_{K,h}(i, j) = \prod_{k=j}^{i-1} \{T_K((k+1)h, kh) + h\Phi_h(k)\},$$

$$(4.8) \quad T_{V,h}(i, j) = \prod_{k=j}^{i-1} \{T_V((k+1)h, kh) + h\Psi_h(k)\},$$

for $0 \leq j < i < \infty$, where

$$\Phi_h(k) = \frac{1}{h} \left(\int_{kh}^{(k+1)h} T((k+1)h, \eta)B(\eta)[K(kh) - K(\eta)T_V(\eta, kh)]d\eta \right)$$

and

$$\Psi_h(k) = \left(V(kh)Q_h(k)^{1/2} - \frac{1}{h} \int_{kh}^{(k+1)h} T((k+1)h, \eta)V(\eta)Q(\eta)^{1/2}T_V(\eta, kh)d\eta \right),$$

for $k \geq 0$. Let $0 < \omega \leq \alpha$ and define $\hat{T}_{K,h}(i, j) = \exp\{\omega(i-j)h\}T_{K,h}(i, j)$, $0 \leq j \leq i < \infty$ and $\hat{T}_K(t, s) = \exp\{\omega(t-s)\}T_K(t, s)$, $0 \leq s \leq t < \infty$. We define \hat{T}_V and $\hat{T}_{V,h}$ analogously. It is not difficult to verify that

$$\|\hat{T}_K(t, s)\|_{L(H)} \leq M, \quad \|\hat{T}_V(t, s)\|_{L(H)} \leq M.$$

Multiplying both sides of (4.7) and (4.8) by $\exp\{\omega(i-j)h\}$, and rewriting these equations in a variation of constants form, we obtain

$$\begin{aligned} \hat{T}_{K,h}(i, j) &= \hat{T}_K(ih, jh) + h \sum_{k=j}^{i-1} \hat{T}_K(ih, (k+1)h) e^{\omega h} \Phi_h(k) \hat{T}_{K,h}(k, j), \\ \hat{T}_{V,h}(i, j) &= \hat{T}_V(ih, jh) + h \sum_{k=j}^{i-1} \hat{T}_V(ih, (k+1)h) e^{\omega h} \Psi_h(k) \hat{T}_{V,h}(k, j). \end{aligned}$$

If there exists a constant $h_0 > 0$ such that for all $h \leq h_0$, $\exp\{\omega h\} \|\Phi_h(k)\|_{L(H)} \leq \omega/2M$ and $\exp\{\omega h\} \|\Psi_h(k)\|_{L(H)} \leq \omega/2M$, then

$$\begin{aligned} \|\hat{T}_{K,h}(i, j)\|_{L(H)} &\leq M + h \sum_{k=j}^{i-1} M \cdot \frac{\omega}{2M} \|\hat{T}_{K,h}(k, j)\|_{L(H)}, \\ \|\hat{T}_{V,h}(i, j)\|_{L(H)} &\leq M + h \sum_{k=j}^{i-1} M \cdot \frac{\omega}{2M} \|\hat{T}_{V,h}(k, j)\|_{L(H)}. \end{aligned}$$

The discrete Gronwall inequality then yields

$$\begin{aligned} \|\hat{T}_{K,h}(i, j)\|_{L(H)} &\leq M e^{\omega(i-j)h/2}, \\ \|\hat{T}_{V,h}(i, j)\|_{L(H)} &\leq M e^{\omega(i-j)h/2}. \end{aligned}$$

Therefore, $T_{K,h}$ and $T_{V,h}$ are uniformly exponentially stable for all $0 < h \leq h_0$.

It is not difficult to see that for each $k \geq 0$, $\Phi_h(k)$ and $\Psi_h(k)$ converge strongly to zero as h tends toward zero. We can obtain convergence in norm if the rank of the operator valued functions $\Phi_h(k)$ and $\Psi_h(k)$ is finite.

DEFINITION 4.1. (Finite rank operator-valued function) Let X and Y be Hilbert spaces with inner products $\langle \cdot, \cdot \rangle_X$ and $\langle \cdot, \cdot \rangle_Y$, respectively. An operator-valued function $W(\cdot) : [0, \infty) \mapsto L(X, Y)$ is said to be continuous and to have finite rank, if there exist continuous vector-valued functions $f_k(\cdot) : [0, \infty) \mapsto X$ and $g_k(\cdot) : [0, \infty) \mapsto Y$, $k = 1, \dots, n$ with $n < \infty$, such that for all $x \in X$,

$$W(t)x = \sum_{k=1}^n \langle f_k(t), x \rangle_Y g_k(t).$$

We define the following condition.

(F) Finite rank stabilizability and detectability. There exist finite rank continuous operator-valued functions $K(\cdot), V(\cdot)$ such that the perturbed evolution systems T_K and T_V are exponentially stable.

LEMMA 4.1. Suppose that conditions (A1)–(A4) hold. If the finite rank condition (F) is satisfied, then on any finite subinterval of $[0, \infty)$, the operator-valued functions $\hat{\Phi}_h$, and $\hat{\Psi}_h$ constructed from Φ_h and Ψ_h in the usual manner, converge uniformly to zero in the uniform operator norm as h tends toward zero.

Proof. We consider $\tilde{\Psi}_h$ only; the argument for $\tilde{\Phi}_h$ is analogous. Using the finite rank condition, we write

$$V(t)\phi = \sum_{k=1}^n \langle f_k(t), \phi \rangle_H g_k(t),$$

with f_k and g_k continuous for $k = 1, \dots, n$. It follows that

$$V(ih)Q_h(i)^{1/2}\phi = \sum_{k=1}^n \langle Q_h(i)^{1/2}f_k(ih), \phi \rangle_H g_k(ih),$$

for $i \geq 0$, and

$$T(t, \eta)V(\eta)Q(\eta)^{1/2}T_V(\eta, s)\phi = \sum_{k=1}^n \langle T_V(\eta, s)^*Q(\eta)^{1/2}f_k(\eta), \phi \rangle_H T(t, \eta)g_k(\eta),$$

for $0 \leq s \leq \eta \leq t < \infty$. Therefore, we have

$$\begin{aligned} \tilde{\Psi}_h(t)\phi &= V([t/h]h)\tilde{Q}_h(t)^{1/2}\phi \\ &\quad - \frac{1}{h} \int_{[t/h]h}^{([t/h]+1)h} T([t/h]+1)h, \eta V(\eta)Q(\eta)^{1/2}T_V(\eta, [t/h]h)\phi d\eta \\ &= \frac{1}{h} \sum_{k=1}^n \int_{[t/h]h}^{([t/h]+1)h} \left\{ \langle \tilde{Q}_h(t)^{1/2}f_k([t/h]h), \phi \rangle_H g_k([t/h]h) \right. \\ &\quad \left. - \langle T_V(\eta, [t/h]h)^*Q(\eta)^{1/2}f_k(\eta), \phi \rangle_H T([t/h]+1)h, \eta g_k(\eta) \right\} d\eta. \end{aligned}$$

By adding and subtracting the term $\langle T_V(\eta, [t/h]h)^*Q(\eta)^{1/2}f_k(\eta), \phi \rangle_H g_k([t/h]h)$ under each of the above integral signs, and using the Schwartz inequality, we obtain the following estimate:

$$\begin{aligned} \|\tilde{\Psi}_h(t)\phi\|_H &\leq \frac{1}{h} \sum_{k=1}^n \int_{[t/h]h}^{([t/h]+1)h} \left\{ w_h(t, \eta) \|g([t/h]h)\|_H \right. \\ &\quad \left. + v_h(t, \eta) \|T_V(\eta, [t/h]h)^*Q(\eta)^{1/2}f_k(\eta)\|_H \right\} \|\phi\|_H d\eta \\ &= \frac{1}{h} \sum_{k=1}^n \int_{[t/h]h}^{([t/h]+1)h} u_h(t, \eta) \|\phi\|_H d\eta, \end{aligned}$$

where

$$\begin{aligned} w_h(t, \eta) &= \|\tilde{Q}_h(t)^{1/2}f_k([t/h]h) - T_V(\eta, [t/h]h)^*Q(\eta)^{1/2}f_k(\eta)\|_H, \\ v_h(t, \eta) &= \|g([t/h]h) - T([t/h]+1)h, \eta g(\eta)\|_H, \quad \text{and} \\ u_h(t, \eta) &= w_h(t, \eta) \|g([t/h]h)\|_H + v_h(t, \eta) \|T_V(\eta, [t/h]h)^*Q(\eta)^{1/2}f_k(\eta)\|_H. \end{aligned}$$

Since the functions f_k and g_k are continuous on any bounded subinterval $[a, b]$ of $[0, \infty)$, and for any $\epsilon > 0$, there exists $\delta > 0$ such that $\|f_k(t) - f_k(s)\|_H \leq \epsilon$ and $\|g_k(t) - g_k(s)\|_H \leq \epsilon$ for all $t, s \in [a, b]$ with $|t-s| \leq \delta$ and $k = 1, \dots, n$. Then, the boundedness of the operator families T, T_V, V, Q_h, Q and the uniform strong convergence of \tilde{Q}_h to Q implies that for any bounded subinterval $[a, b]$ of $[0, \infty)$, and for any given constant

$\epsilon > 0$, we can find $h_0 > 0$ such that for all $0 < h \leq h_0$ and $t \in [a, b]$, the functions $u_h(t, \eta) \leq \epsilon$ for $\eta \in [t, t + h]$ and $t \in [a, b]$. Consequently, $\|\tilde{\Psi}_h(t)\|_{L(H)} \leq \epsilon$ for all $t \in [a, b]$. \square

We can extend the uniform convergence on finite-time intervals to uniform convergence on the infinite-time interval by assuming certain periodicity (in particular, time invariance) of the evolution system T and the operator-valued functions B, Q, \tilde{Q}_h, K , and V . In fact, the periodicity assumption implies that $\tilde{\Phi}_h, \tilde{\Psi}_h$ are also periodic functions of time. Thus, using Lemma 4.1, we obtain the following theorem.

THEOREM 4.1. *Assume that the evolution system T and the operator-valued functions B, Q, R are strongly continuous and periodic with the same period θ . Suppose further that the periodicity of Q is preserved by \tilde{Q}_h for the sampled-time systems. If the finite rank condition (F) holds for some θ -periodic functions K and V , then the discretization defined in (4.3) and (4.4) generates uniformly stabilizable and uniformly detectable sampled control systems for sampling periods h with $0 < h \leq h_0$ for some constant $h_0 > 0$.*

The periodicity assumption is trivially satisfied in a large number of practical examples; in particular, it is satisfied for all time-invariant systems. However, the finite rank assumption says, in essence, that only a finite number of modes of the state vector are unstable in the absence of control. Indeed, in the case of evolution systems corresponding to a hyperbolic partial differential equation, there exists examples in which if the finite rank condition is not satisfied, all sampled systems are not stabilizable, even though the continuous-time control system is stabilizable. However, stabilizable systems whose open-loop dynamics are described by compact, analytic, or differentiable semigroups and whose unstable manifold is finite-dimensional, can be stabilized via finite rank feedback. These results have been reported in [RW] and [RW2]. Thus, the arguments presented here are not as restrictive as they seem. For other discretization schemes, the uniform stabilizability and uniform detectability of the generated sampled systems remains, in most cases, an open question.

5. Examples and numerical results. In this section, we present and briefly discuss some of our numerical findings, which serve to illustrate our convergence results in the context of a variety of distributed parameter control systems. In particular, we consider the infinite-horizon optimal control or regulation of a heat or diffusion equation, a delay or hereditary system, and a flexible structure in the form of a cantilevered Voigt–Kelvin viscoelastic beam with tip mass.

In all of the examples to follow, we consider time-invariant systems only, and obtain the discrete- or sampled-time operators from the corresponding continuous time operators via $T_h = T(h)$, $B_h = h^{-1} \int_0^h T(t)Bdt$, $Q_h = Q$, and $R_h = R$, for $h > 0$ (i.e., via zero-order hold sampling). To solve the resulting infinite-dimensional continuous and discrete-time LQR problems, we introduced some form of state discretization (i.e., either modal or spline-based Ritz–Galerkin techniques), which were known to yield convergence in the closed-loop problem. By choosing the state discretization sufficiently fine, we could assume that we obtained a reasonably accurate finite-dimensional approximation to the solution of the infinite-dimensional LQR problems.

The resulting finite-dimensional continuous- and discrete-time LQR problems (more precisely, the matrix algebraic Riccati equations) were solved using either eigenvector (in the continuous-time case, also known as Potter's method; see [KS]) or Schur vector (for the discrete-time problems; see [PLS]) decomposition of the Hamiltonian matrix. All computations for the first two examples were carried out on an IBM PC

AT. The flexible structure problem was solved on an IBM3090, although it, too, could have been solved on a personal computer.

In each of the examples below, the control systems are time invariant and the control space U is finite-dimensional. In fact, $U = R$. Thus, the optimal feedback gains F and F_h are elements in $L(H, R)$. That is, they are bounded linear functionals on H . Consequently, they admit representors, respectively, f and f_h , in H with $F\varphi = \langle f, \varphi \rangle_H$ and $F_h\varphi = \langle f_h, \varphi \rangle_H$, for $\varphi \in H$. The elements f and f_h in H are referred to as the optimal continuous- or discrete-time functional feedback control gains. The finite dimensionality of the control space U also implies the uniform stabilizability of the sampled systems when the continuous-time systems are stabilizable (recall Theorem 4.1). Our convergence result implies that $\lim_{h \rightarrow 0^+} F_h\varphi = F\varphi$ for $\varphi \in H$. Note that when U is finite-dimensional, this is equivalent to $\lim_{h \rightarrow 0^+} F_h = F$ in the uniform norm topology on $L(H, U)$ and $\lim_{h \rightarrow 0^+} f_h = f$ in H . It is this latter convergence of the functional gains that we shall exhibit in our plots below.

Example 5.1. We consider the scalar or one-dimensional heat or diffusion control system

$$\frac{\partial}{\partial t}x(t, \eta) = a \frac{\partial^2}{\partial \eta^2}x(t, \eta) + b\chi_{[\epsilon_1, \epsilon_2]}(\eta)u(t), \quad 0 < \eta < 1, t > 0,$$

with the Dirichlet boundary conditions

$$x(t, 0) = x(t, 1) = 0, \quad t > 0,$$

at $\eta = 0$ and $\eta = 1$, where $a > 0, b \in R, 0 \leq \epsilon_1 < \epsilon_2 \leq 1$, and χ_s denotes the characteristic function on the set S . We take the performance index to be

$$J(u) = \int_0^\infty \left\{ \int_0^1 qx(t, \eta)^2 d\eta + ru(t)^2 \right\} dt,$$

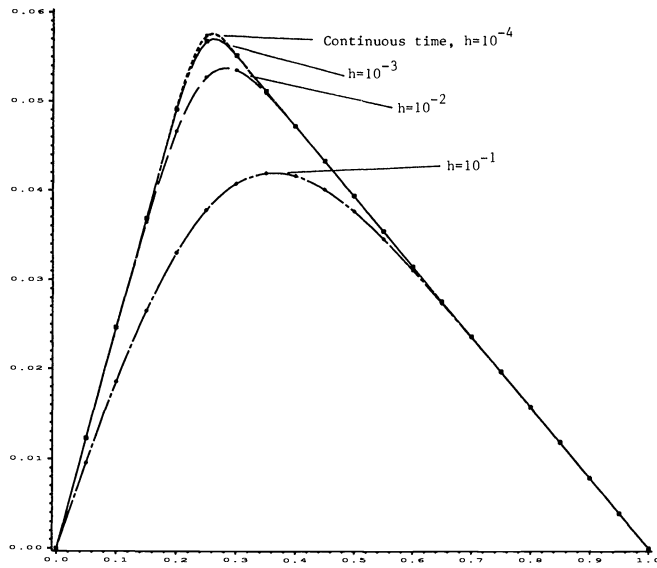
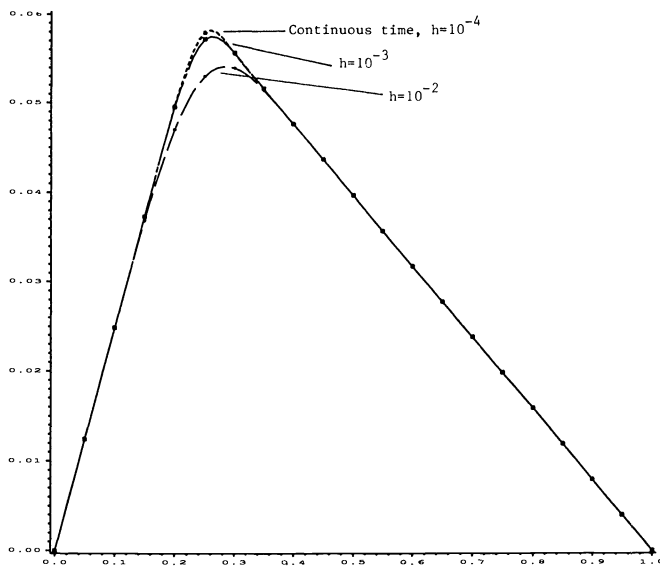
with $q \geq 0$ and $r > 0$.

In this case, we have $H = L_2(0, 1), U = R, A : \text{Dom}(A) \subset H \mapsto H$ given by

$$A\varphi = aD^2\varphi \text{ for } \varphi \in \text{Dom}(A) = H^2(0, 1) \cap H_0^1(0, 1),$$

$B \in L(R, H)$ given by $(Bv)(\eta) = b\chi_{[\epsilon_1, \epsilon_2]}(\eta)v, 0 < \eta < 1, v \in R, Q \in L(H)$ given by $Q = qI$, and $R \in L(U)$ given by $R = rI$, where I denotes the identity map on R . We note that $\{T(t) : t \geq 0\}$, the semigroup of bounded linear operators on H with infinitesimal generator A , is parabolic and uniformly exponentially stable. Thus the continuous-time pairs, $\{A, B\}$ and $\{Q, A\}$ are trivially stabilizable and detectable, and the discrete-time pairs, $\{T_h, B_h\}$ and $\{Q_h, T_h\}$ are uniformly stabilizable and detectable as well.

Setting $a = 0.1, b = 1.0, q = 1.0, r = 1.0, \epsilon_1 = 0.21$, and $\epsilon_2 = 0.275$, we obtained the plot of the functional gains f and f_h in $L_2(0, 1)$, for various values of $h > 0$, given in Figs. 5.1 and 5.2. Those in Fig. 5.1 were obtained via a modal (i.e., $\sin(k\pi x), k = 1, 2, \dots, N$) state discretization with $N = 20$ modal elements. For the gains in Fig. 5.2, we used linear B-spline elements (i.e., "hat" functions) defined with respect to a uniform partition of $[0, 1]$ into $N = 20$ subintervals of equal length. Convergence of these state approximations and the corresponding closed-loop solutions to the control problem is well known (see, for example, [G], [GR], and [R]).

FIG. 5.1. *Functional gains for heat equation with modal approximation*FIG. 5.2. *Functional gains for heat equation with spline approximation*

Example 5.2. In this example, we consider the scalar, single input hereditary control system

$$(5.1) \quad \dot{x}(t) = a_0 x(t) + a_1 x(t-1) + bu(t),$$

where $a_0, a_1, b \in R$. We take the performance index to be

$$J(u) = \int_0^\infty \{qx^2(t) + ru^2(t)\} dt,$$

TABLE 5.1
Head gains for hereditary system.

Sampling period h	Head gain f^0
10^{-1}	3.76185
10^{-2}	4.35007
10^{-3}	4.41577
10^{-4}	4.42241
10^{-5}	4.42308
10^{-6}	4.42314
Continuous time	4.42315

with $q \geq 0$ and $r > 0$.

The abstract Hilbert space formulation for linear hereditary control systems is well known (see, for example, [BB]). We let $H = R \times L_2(-1, 0)$, $U = R$ and set $A : \text{Dom}(A) \subset H \mapsto H$ to be $A(\eta, \varphi) = (a_0\eta + a_1\varphi(-1), D\varphi)$ for $(\eta, \varphi) \in \text{Dom}(A) = \{(\xi, \psi) \in H : \psi \in H^1(-1, 0), \xi = \psi(0)\}$. The operator A is the infinitesimal generator of the C_0 -semigroup of bounded linear operators on H , $\{T(t) : t \geq 0\}$, given by $T(t)(\eta, \varphi) = (x(t), x_t)$ where x is the solution to (5.1) with $u \equiv 0$ and corresponding to the initial data $x(0) = \eta$, $x(\theta) = \varphi(\theta)$, $-1 \leq \theta \leq 0$, and $x_t \in L_2(-1, 0)$ is the past history of x from t back to $t - 1$. That is $x_t(\theta) = x(t + \theta)$, $-1 \leq \theta \leq 0$. We let $B \in L(R, H)$, $Q \in L(H)$, and $R \in L(U)$ be given by $Bv = (bv, 0)$, $Q(\eta, \varphi) = (q\eta, 0)$, and $Rv = rv$, respectively.

To solve both the continuous- and discrete-time LQR problems we employed a piecewise constant/linear spline hybrid finite element scheme developed by Ito and Kappel in [IK]. Setting $a_0 = a_1 = b = q = r = 1$, and with a state discretization level in the Ito–Kappel scheme taken to be $N = 20$, we obtained the $R \times L_2(-1, 0)$ functional gains, $f = (f^0, f^1)$ and $f_h = (f_h^0, f_h^1)$ for various values of $h > 0$, tabulated and plotted in Table 5.1 and Fig. 5.3.

We note that for this choice of the parameters a_0 and a_1 , the open-loop system has an eigenvalue with positive real part. Consequently, system (5.1) is open-loop unstable. It is not difficult to argue that the pairs $\{A, B\}$ and $\{Q, A\}$ are, respectively, stabilizable and detectable. Also, since the operators B and Q are of finite rank, there exists $h_0 > 0$ such that for all sampling periods $h \leq h_0$, the sampled control systems are uniformly stabilizable and detectable in h .

Example 5.3. We consider the control of the small amplitude transverse vibration of a cantilevered Voigt–Kelvin viscoelastic beam with tip-mass. The relevant dynamics are described by the hybrid system of ordinary and partial differential equations

$$\begin{aligned} \rho \frac{\partial^2}{\partial t^2} x(t, \eta) + cI \frac{\partial^5}{\partial \eta^4 \partial t} x(t, \eta) + EI \frac{\partial^4}{\partial \eta^4} x(t, \eta) &= 0, \quad \eta \in (0, 1), \\ m \frac{\partial^2}{\partial t^2} x(t, 1) - cI \frac{\partial^4}{\partial \eta^3 \partial t} x(t, 1) - EI \frac{\partial^3}{\partial \eta^3} x(t, 1) &= bu(t), \end{aligned}$$

for $t > 0$, the essential (or stable) boundary conditions at $\eta = 0$

$$x(t, 0) = 0, \quad \frac{\partial}{\partial \eta} x(t, 0) = 0, \quad t > 0,$$

and the natural (or unstable) boundary condition at $\eta = 1$,

$$cI \frac{\partial^3}{\partial \eta^2 \partial t} x(t, 1) - EI \frac{\partial^2}{\partial \eta^2} x(t, 1) = 0, \quad t > 0.$$

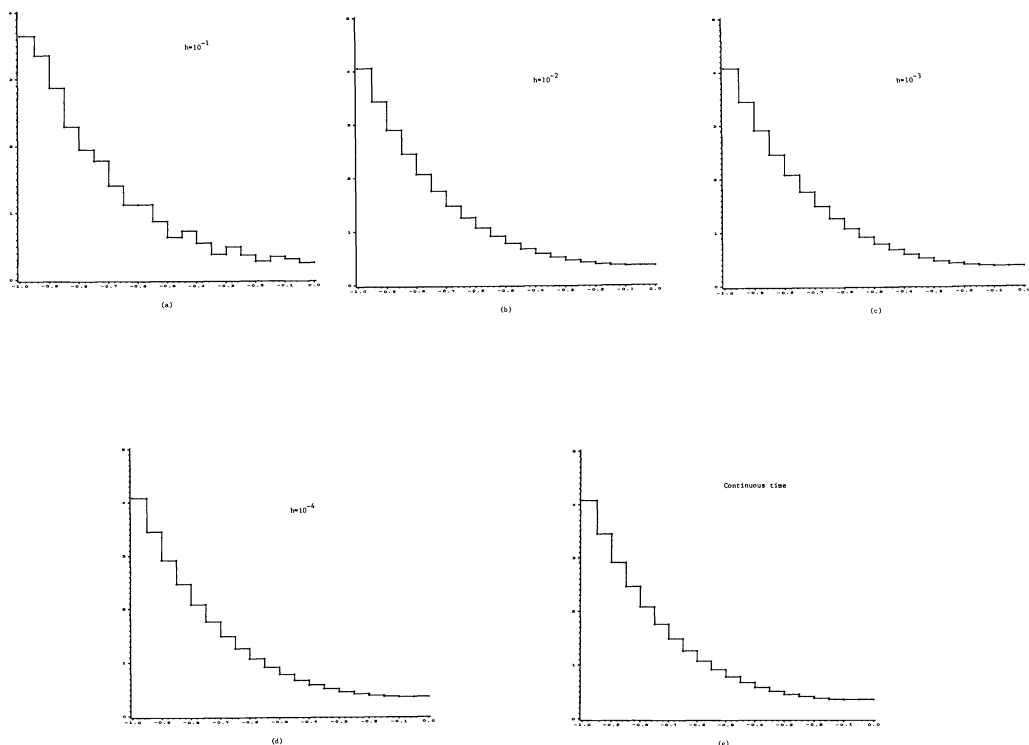


FIG. 5.3. *Functional gains for delay equation:* (a) $h=0.1$, (b) $h=0.01$, (c) $h=0.001$, (d) $h=0.0001$, (e) *continuous time*.

In the above equations, $\rho > 0$ is the linear mass density of the beam, $I > 0$ is the beam's cross-sectional moment of inertia, $c > 0$ is the viscosity coefficient, $E > 0$ is Young's modulus, $m > 0$ is the mass of the tip mass, and $b \in \mathbb{R}$ is a constant.

We take an energy-based performance index

$$J(u) = \int_0^\infty \left\{ \frac{1}{2} EI \int_0^1 \left(\frac{\partial^2}{\partial \eta^2} x(t, \eta) \right)^2 d\eta + \frac{1}{2} m \left(\frac{\partial}{\partial t} x(t, 1) \right)^2 + \int_0^1 \frac{1}{2} \rho \left(\frac{\partial}{\partial t} x(t, \eta) \right)^2 d\eta + ru(t)^2 \right\} dt.$$

Once again, the abstract Hilbert space formulation of this problem is standard. We let $H = H_L^2(0, 1) \times \mathbb{R} \times L_2(0, 1)$, where $H_L^2(0, 1) = \{\varphi \in H^2(0, 1) : \varphi(0) = D\varphi(0) = 0\}$, and endow H with the energy inner product

$$\begin{aligned} \langle (\varphi_1, \eta_1, \psi_1), (\varphi_2, \eta_2, \psi_2) \rangle_H &= EI \int_0^1 D^2 \varphi_1 D^2 \varphi_2 \\ &\quad + m \eta_1 \eta_2 + \rho \int_0^1 \psi_1 \psi_2. \end{aligned}$$

TABLE 5.2
Tip gains for beam equation.

Sampling period h	Tip Gain, f^1
1.000	0.12181
0.500	0.12003
0.010	0.11798
0.005	0.11796
0.001	0.11794
Continuous time	0.11793

The operator $A : \text{Dom}(A) \subset H \mapsto H$ is given by $A(\varphi, \eta, \psi) = (\psi, cID^3\psi(1) + EID^3\varphi(1), -cID^4\psi - EID^4\varphi)$ for $(\varphi, \eta, \psi) \in \text{Dom}(A) = \{(\varphi, \eta, \psi) \in H : \psi \in H_L^2(0, 1), \eta = \psi(1), cID^2\psi + EID^2\varphi \in H^2(0, 1), cID^2\psi(1) + EID^2\varphi(1) = 0\}$. We take $U = R$ and define $B \in L(R, H)$ by $Bv = (0, bv, 0)$. We let $Q \in L(H)$ and $R \in L(U)$ be given by $Q = (1/2)I_H$ and $R = rI_U$, where I_H , and I_U denote, respectively, the identity operators on H and U .

It can be shown (see [GA]) that A is the infinitesimal generator of a uniformly exponentially stable analytic semigroup. Thus, once again, stabilizability and detectability for the continuous-time problems trivially follows, as does the uniform stabilizability and detectability for the discrete-time problems.

We employed a standard cubic spline based Ritz–Galerkin finite element scheme to approximate or finite-dimensionalize the continuous- and discrete-time LQR problems (see [GA], [GR]). Setting $\rho = 0.1$, $EI = 1.3333 \times 10^{-4}$, $cI = 1.3333 \times 10^{-7}$, $m = 1$, $b = 1$, $q = 1$, and $r = 1$ and with $N = 9$ cubic spline elements, we obtained the functional gains $f = (f^0, f^1, f^2)$, $f_h = (f_h^0, f_h^1, f_h^2) \in H$ exhibited in Table 5.2 and Fig. 5.4.

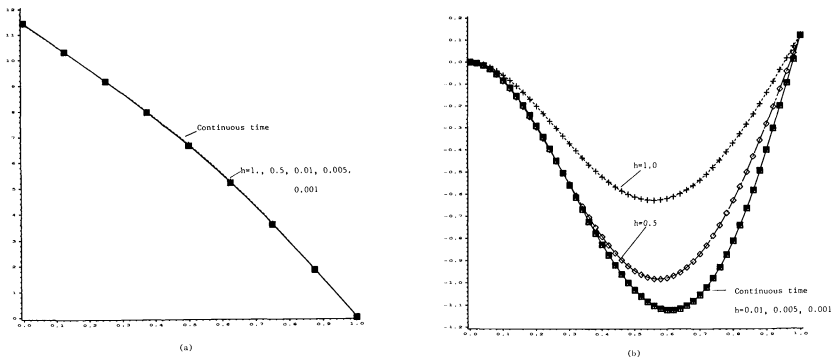


FIG. 5.4. Functional gains for beam equation. (a) displacement, we plot $D^2 f^0$ to exhibit the H^2 -convergence; (b) velocity.

6. Summary and concluding remarks. We have investigated and established the convergence of solutions to discrete- or sampled-time linear quadratic regulator

problems and the associated Riccati equations for infinite-dimensional systems to the solutions to the corresponding continuous-time problem and associated Riccati equation, as the length of the sampling interval tends toward zero. We have considered both the finite- and infinite-time horizon problems and carried out numerical studies involving a variety of distributed parameter control systems to observe how well our theoretical results predict what actually takes place in practice. In the context of the finite-time horizon problem, the assumption of strong continuity on the operators that define the control system and performance index, together with a stability and consistency hypothesis on the sampling scheme, are sufficient to establish the strong convergence of the Riccati operators, feedback gains, optimal control laws, and optimal trajectories, with some degree of uniformity in time over the compact interval of interest. For the infinite-time horizon problem, we require the additional assumption of stabilizability and detectability, uniformly with respect to the length of the sampling interval. We have shown that this condition can be verified when zero-order hold sampling is employed and the continuous-time system is stabilizable and detectable by finite rank feedback. We have shown that this can be done for certain classes of systems and, in particular, stabilizable systems whose open-loop dynamics are described by compact, analytic, or differentiable semigroups, and whose unstable manifold is finite-dimensional. These results are reported on elsewhere (see [RW] and [RW2]).

Several interesting questions related to the results we have presented here remain open. For example, the interrelation between stabilizability/detectability for the continuous- and sampled-time systems in a more general setting and under more general sampling schemes (A-Stable Padé, for example) requires further study. Also, convergence under simultaneous and independent state (space) discretization (i.e., finite difference or finite element approximation) and temporal sampling should be investigated. It would not be difficult to extend the results presented here to handle certain "coupled" state and time discretizations. Finally, a study similar to the present one could be carried out for the LQG estimator and compensator problems. We have not as of yet looked at these problems, but suspect that similar results to those given above could be obtained.

REFERENCES

- [A] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
- [BB] H. T. BANKS AND J. A. BURNS, *Hereditary control problems: Numerical methods based on averaging approximations*, SIAM J. Control Optim., 16 (1978), pp. 169–208.
- [BKS] H. T. BANKS, S. L. KEELING, AND R. J. SILCOX, *Optimal control techniques for active noise suppression*, in Proc. 27th IEEE Conference on Decision and Control, Austin, TX, Dec. 7–9, 1988, pp. 2006–2011.
- [BKSW] H. T. BANKS, S. L. KEELING, R. J. SILCOX, AND C. WANG, *Linear quadratic tracking problems in Hilbert Space: Application to optimal active noise suppression*, in IFAC Proc. 5th Symposium on Control of Distributed Parameter Systems, Perpignan, France, June 26–29, 1989, pp. 17–22.
- [BK] H. T. BANKS AND K. KUNISCH, *The linear regulator problem for parabolic systems*, SIAM J. Control Optim., 22 (1984), pp. 684–698.
- [BW] H. T. BANKS AND C. WANG, *Optimal feedback control of infinite-dimensional parabolic evolution systems: approximation techniques*, SIAM J. Control Optim., 27 (1989), pp. 1182–1219.
- [C] G. CHEN, *Energy decay estimates and exact boundary value controllability for wave equation in a bounded domain*, J. Math. Pures Appl., 58 (1979), pp. 249–273.
- [D] R. DATKO, *Uniform asymptotic stability of evolution processes in a Banach space*, SIAM J. Math. Anal., 3 (1972), pp. 428–445.

- [DI] G. DA PRATO AND A. ICHIKAWA, *Quadratic control for linear time varying systems*, SIAM J. Control Optim., 28 (1990), pp.359–381.
- [G] J. S. GIBSON, *The Riccati integral equations for optimal control problems on Hilbert spaces*, SIAM J. Control Optim., 17 (1979), pp. 537–565.
- [GA] J. S. GIBSON AND A. ADAMIAN, *Approximation theory for LQG optimal control of flexible structures*, ICASE Report No.88-48, Institute for Computer Applications in Science and Engineering, Hampton, VA, 1988; SIAM J. Control Optim., 29 (1991), pp. 1–37.
- [GR] J. S. GIBSON AND I. G. ROSEN, *Numerical approximation for the infinite dimensional discrete time optimal linear quadratic regulator problem*, SIAM J. Control Optim., 26 (1988), pp.428–451.
- [HH] W. W. HAGER AND L. L. HOROWITZ, *Convergence and stability properties of the discrete Riccati Operator equation and the associated optimal control and filtering problems*, SIAM J. Control Optim., 14 (1976), pp. 295–312.
- [HK] R. HERSH AND T. KATO, *High-accuracy stable difference schemes for well-posed initial-value problems*, SIAM J. Numer. Anal. 16 (1979), pp. 670–682.
- [IK] K. ITO AND F. KAPPEL, *A uniformly differentiable approximation scheme for delay systems using splines*, Appl. Math and Opt., 23 (1991), pp. 217–262.
- [K] T. KATO, *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, Berlin, 1976.
- [KS] H. KWAKERNAAK AND R. SIVAN, *Linear Optimal Control Systems*, Wiley-Interscience, New York, 1972.
- [L] J. LAGNESE, *Decay of solutions of wave equations in a bounded region with boundary dissipation*, J. Differential Equations, 50 (1983), pp.163–182.
- [LCB] K. Y. LEE, S. CHOW, AND R. O. BARR, *On the control of discrete-time distributed parameter systems*, SIAM J. Control Optim., 10 (1972), pp. 361–376.
- [Le] F. L. LEWIS, *Optimal Estimation with an Introduction to Stochastic Control Theory*, Wiley-Interscience, New York, 1986.
- [PLS] T. PAPPAS, A.J. LAUB, AND N. R. SANDELL, JR., *On the numerical solution of the discrete-time algebraic Riccati equation*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 631–641.
- [RM] R. D. RICHTMYER AND K. W. MORTON, *Difference Methods for Initial Value Problems*, Wiley-Interscience, New York, 1967.
- [R] I. G. ROSEN, *Optimal discrete-time LQR problems for parabolic systems with unbounded input-approximation and convergence*, Control Theory Adv. Tech., 5 (1989), pp.277–300.
- [RW] I. G. ROSEN AND C. WANG, *Finite rank stabilizability and detectability preservation under sampling for distributed parameter systems*, in Proc. 29th IEEE Conference on Decision and Control, Dec. 5–7, 1990, Honolulu, HI, pp. 375–376.
- [RW1] ———, *On the continuous dependence with respect to sampling of the linear quadratic regulator problem for distributed parameter systems*, ICASE Report No. 90-23, Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, VA, March, 1990.
- [RW2] ———, *On stabilizability and sampling for infinite dimensional systems*, IEEE Trans. Auto. Control, to appear.
- [W] C. WANG, *Approximation methods for linear quadratic regulator problems with nonautonomous periodic parabolic systems*, Ph.D thesis, Brown University, Providence, RI, May, 1988.
- [Z] J. ZABCZYK, *Remarks on the control of discrete-time distributed parameter systems*, SIAM J. Control Optim., 12 (1974), pp. 721–735.

ON THE PRINCIPLE OF SMOOTH FIT FOR A CLASS OF SINGULAR STOCHASTIC CONTROL PROBLEMS FOR DIFFUSIONS*

JIN MA[†]

Abstract. This paper considers the principle of smooth fit for a class of one-dimensional singular stochastic control problems allowing the system to be of nonlinear diffusion type. The existence and the uniqueness of a convex C^2 -solution to the corresponding variational inequality are obtained. It is proved that this solution gives the value function of the control problem, and the optimal control process is constructed. As an example of the degenerate case, it is proved that the conclusion is also true for linear systems, and the explicit formula for the smooth fit points is derived.

Key words. singular stochastic control, principle of smooth fit, variational inequality, free boundary problem, diffusion with reflections

AMS(MOS) subject classifications. 93E20, 49A10, 49A60

1. Introduction. Let $(\Omega, \mathcal{F}, P; \mathcal{F}_t)$ be a complete probability space with filtration $\{\mathcal{F}_t\}$, which is assumed to be right-continuous, and \mathcal{F}_0 contains all the P -null sets in \mathcal{F} . We assume that a one-dimensional standard Brownian motion $W = \{W(t) : t \geq 0\}$ with respect to $\{\mathcal{F}_t\}$ is given on this probability space.

Consider the system described by the stochastic differential equation

$$(1.1) \quad dX(t) = a(X(t))dt + \sigma(X(t))dW(t) + d\xi(t), \quad X(0) = x,$$

or, equivalently, by the stochastic integral equation

$$(1.2) \quad X(t) = x + \int_0^t a(X(s))ds + \int_0^t \sigma(X(s))dW(s) + \xi(t),$$

where $\xi = \{\xi(t) : t \geq 0\}$ is a *left-continuous, $\{\mathcal{F}_t\}$ -adapted process with locally bounded variation paths*. The process ξ is to be chosen by the decision maker as the *control process*, and the objective is to minimize the following cost function:

$$(1.3) \quad V_\xi(x) = E \int_{[0, \infty)} e^{-\alpha t} [cd\check{\xi}(t) + h(X(t))dt],$$

where $\check{\xi} = \{\check{\xi}(t) : t \geq 0\}$ is the *total variation process* of ξ ; the constant $\alpha > 0$ is called the *discount factor*; h is a nonnegative, strictly convex, C^2 -function; and $c > 0$.

Problems of similar type have been studied by many authors (cf. [1], [2], [6], [8]–[10], [12]–[14]). In the case when $c = 0$, $h(x) = x^2$, $a(x) \equiv 0$, $\sigma(x) \equiv 1$, the problem was solved explicitly by Beněš, Shepp, and Witsenhausen [1] under the constraints that either ξ has bounded derivatives (bounded velocity follower problem) or it has bounded total variation (finite-fuel follower problem). Under the same setting but without the extra restriction on ξ , and allowing h to be a general strictly convex function and $c = 1$, the result was generalized by Karatzas [8]. Almost simultaneously, Harrison and Taksar [6] treated the case with a more general cost function but restricted (compact) state space and, also, they assumed the drift and the diffusion coefficients to be constants.

* Received by the editors December 19, 1990; accepted for publication (in revised form) May 24, 1991. This work is a part of the author's Ph.D. dissertation at the University of Minnesota.

[†] Department of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455.

For the case when a, σ are nonconstant, the problem was developed by Menaldi and Robin [9], Chow, Menaldi, and Robin [2], and Shreve, Lehoczky, and Gaver [13], among others. In [9], however, no control entered the cost function explicitly (i.e., $c = 0$), and, for the convex case (i.e., h is convex), the result there was only valid when a and σ are constants. In [2] the horizon was assumed to be finite, and only the monotone follower problem (i.e., ξ is monotone) was considered. We note that the bounded variation control problem (optimal correction problem) was considered there only when some special symmetric conditions were satisfied by h , so that the problem could be reduced to the monotone follower problem. In general, however, these conditions are not satisfied in our setting. Finally, in [13], it was essentially the homogeneous problem (i.e., $h(\cdot) \equiv 0$), so the problem is quite different from ours. We note that, for the convexity of the value function, all the above work required the coefficients of the system to be constant or linear (in spatial variables), so that the convexity of the function h would imply the convexity of value function immediately. However, this requirement is not satisfied, in general, in our setting.

The problem is also studied for a higher-dimension case by Soner and Shreve [14] and Menaldi and Taksar [10]; some regularity results for the free boundary, as well as the convexity of the value function, were obtained. However, the difficulties that arise in higher dimensions seem to restrict the problem only to the case when a, σ are constants.

In this paper, we are interested in the system when $a(x) = ax + b$ and when σ is any nonvanishing, Lipschitz continuous, C^2 -function of linear growth. Under some conditions on the discount factor α and the function σ , we prove that the *principle of smooth fit* always holds in this case. Namely, we prove that there exists a unique *convex C^2 -solution* to the variational inequality that is *linear* outside a certain finite interval (even though the data of the system, e.g., σ , could be nonlinear), which, as was pointed out by Shreve [12], gives the value function and leads to the existence of the optimal policy for such problems. Consequently, the optimal policy can then be chosen to be the proper local times to make the dynamics to be the reflected diffusion on a certain region. Compared to the usual way of treating variational inequalities, our approach is direct and elementary but strongly restricted to the one-dimensional case.

An interesting question then is how this setting includes the linear case, namely, when $\sigma(\cdot)$ is also linear. An immediate problem is that the related ordinary differential equation (ODE) becomes singular at some point (the zero of σ). In §5 we treat this case specifically to get an explicit solution.

The paper is organized as follows. In §2 we give the formulation of the problem and the verification theorems. In §3 we study the ODE related to the H-J-B equation and give some basic results as lemmas for the main theorems. Section 4 is devoted to the main results, and, finally, in §5, we study the linear case, which can also be treated as an example for our setting.

2. Formulation of the problem and the verification theorems. We will henceforth consider the system

$$(2.1) \quad X(t) = x + \int_0^t (aX(s) + b)ds + \int_0^t \sigma(X(s))dW(s) + \xi(t),$$

where a, b are constants, $W(\cdot)$ is a one-dimensional Brownian motion with respect to the filtration $\{\mathcal{F}_t\}$.

As in §1, for a given ξ , the cost function is defined by

$$(2.2) \quad V_{\xi}(x) = E \int_{[0, \infty)} e^{-\alpha t} [cd\check{\xi}(t) + h(X(t))dt].$$

We will assume that $c = 1$ for simplicity.

The value function is defined by

$$(2.3) \quad V^*(x) = \inf_{\xi \in \mathcal{B}} V_{\xi}(x), \quad x \in \mathbf{R},$$

where \mathcal{B} is a class of processes called *admissible controls*, which will be described later.

We make the following basic assumptions:

(A1) The function $\sigma : \mathbf{R} \rightarrow \mathbf{R}$ is of class C^2 such that, for some $K > 0$,

$$(2.4) \quad |\sigma'(x)| + |(\sigma^2(x))''| \leq K, \quad x \in \mathbf{R};$$

$$(2.5) \quad \sigma(x) \neq 0, \quad x \in \mathbf{R}.$$

Clearly, (2.4) implies that $\sigma(\cdot)$ is globally Lipschitz and of linear growth; i.e., for some $K_1 > 0, K_2 > 0$,

$$(2.6) \quad |\sigma(x) - \sigma(y)| \leq K_1|x - y|, \quad x, y \in \mathbf{R},$$

and

$$(2.7) \quad |\sigma(x)| \leq K_2(1 + |x|), \quad x \in \mathbf{R},$$

where the constants K_1, K_2 depend only on K and $\sigma(0)$.

(A2) The function $h : \mathbf{R} \rightarrow [0, \infty)$ is of class C^2 such that, for some k, K_3 with $0 < k < K_3$,

$$(2.8) \quad 0 < k \leq h''(x) \leq K_3, \quad x \in \mathbf{R},$$

and there exists $\bar{x} \in \mathbf{R}$ such that

$$(2.9) \quad (x - \bar{x})h'(x) \geq 0, \quad x \in \mathbf{R}; \quad h'(\bar{x}) = 0.$$

Also, for simplicity, we assume that $\bar{x} = 0$.

(A3) The discount factor $\alpha > 0$ satisfies

$$(2.10) \quad \alpha > \frac{1}{2} \sup_{x \in \mathbf{R}} |(\sigma^2(x))''| + 2|a|.$$

Remark 2.1. (i) Condition (A3) seems to be a little strong, since it actually requires that the discount factor be sufficiently large. This is to compensate for the fact that the coefficients are not constant. In fact, without this assumption, the convexity of the value function, which is essential in the smooth fit technique, may be false. The similar condition has also been used in [2], [9], [13], and others.

(ii) By the definition of the function h , the cost function satisfies $V_{\xi}(x) \geq 0$ for all $x \in \mathbf{R}, \xi \in \mathcal{B}$. Also, (2.8) and (2.9) imply that h is strictly convex and (recall that $\bar{x} = 0$), for any $\delta > 0$, there exist $-\infty < r_1 < 0 < r_2 < \infty$ such that $|h'(x)| < \delta, x \in (r_1, r_2)$, and $|h'(r_1)| = |h'(r_2)| = \delta$.

(iii) Throughout the paper, instead of using constants K_1, K_2, K_3, \dots , we use a generic constant $K > 0$, which may vary line by line if no confusion occurs.

As we mentioned before, $\xi = \{\xi(t) : t \geq 0\}$ is an $\{\mathcal{F}_t\}$ -adapted, left-continuous process such that, for each $\omega \in \Omega$, the path $\xi(\cdot, \omega)$ is of locally bounded variation on $[0, +\infty)$ and $\xi(0) = 0$. We may write ξ in its canonical form $\xi = \xi^+ - \xi^-$ as the difference of two nondecreasing processes ξ^+ and ξ^- with $\xi^\pm(0) = 0$. If we assume that the decomposition is minimal, then the total variation process $\check{\xi}$ can be written as $\check{\xi}(t) = \xi^+(t) + \xi^-(t), t \geq 0$. We denote the totality of such ξ 's by \mathcal{B} (*admissible controls*), and denote, for each $[A, B] \subseteq \mathbf{R}$, $\mathcal{B}_{[A, B]} = \{\xi \in \mathcal{B} : X_\xi(t) \in [A, B] \text{ for } t > 0, \text{ almost surely}\}$. Observe that, for $\xi \in \mathcal{B}_{[A, B]}$, $X_\xi(0) = x$ could be outside the interval $[A, B]$, but, after an initial jump, the trajectories of $X_\xi(\cdot)$ will remain in $[A, B]$, (*P*)-almost surely. It is known that, under our basic assumptions, (2.1) has a (pathwise) unique solution for $t \geq 0$ and every $\xi \in \mathcal{B}(\mathcal{B}_{[A, B]})$ (cf. [5]).

Due to the results for the Brownian motion case (cf. [1], [8]), to get a nontrivial lower bound for the cost functions and the sufficient conditions for a cost function to be optimal, we should seek a convex solution of the following *variational inequality*:

$$(2.11) \quad [\alpha V(x) - \frac{1}{2}\sigma^2(x)V''(x) - (ax + b)V'(x) - h(x)] \vee [|V'(x)| - 1] = 0, \quad x \in \mathbf{R}.$$

The following theorem verifies this fact.

THEOREM 2.1. *Suppose that $V : \mathbf{R} \rightarrow \mathbf{R}$ is a C^2 -function satisfying*

$$(2.12) \quad V''(x) \geq 0, \quad x \in \mathbf{R};$$

$$(2.13) \quad |V'(x)| \leq 1, \quad x \in \mathbf{R};$$

$$(2.14) \quad \alpha V(x) \leq \frac{1}{2}\sigma^2(x)V''(x) + (ax + b)V'(x) + h(x), \quad x \in \mathbf{R};$$

then, under assumptions (A1)–(A3), for all $x \in \mathbf{R}$ and all $\xi \in \mathcal{B}$, we have that $V(x) \leq V_\xi(x)$. Consequently, if there exists a $\xi^* \in \mathcal{B}$ such that $V(x) = V_{\xi^*}(x)$, for all $x \in \mathbf{R}$, then

$$V(x) = V_{\xi^*}(x) = V^*(x), \quad x \in \mathbf{R}.$$

Before proving the theorem, we first give a lemma that may be of independent interest. Soner and Shreve [14, Thm. 3.1] used an easier version to prove their result. We note that their version would suffice for the proof of our theorem as well.

LEMMA 2.2. *Let $\xi \in \mathcal{B}$ and $X(\cdot) = X_\xi^x(\cdot)$ be the corresponding solution of (2.1). If $E\check{\xi}(t+) = o(e^{\alpha t})$ and $E \int_0^\infty e^{-\alpha t} |X(t)|^2 dt < \infty$, then*

$$E|X(t+)| = o(e^{\alpha t}), \quad \text{as } t \rightarrow \infty,$$

where here (and in the following) $o(\rho)$ means $\lim_{\rho} o(\rho)/\rho = 0$.

Proof. Since

$$|X(t+)| \leq |X(t)| + |X(t+) - X(t)| = |X(t)| + |\xi(t+) - \xi(t)| \leq |X(t)| + |\check{\xi}(t+)|,$$

it suffices to prove that $E|X(t)| = o(e^{\alpha t})$, as $t \rightarrow \infty$. By (2.1), we have that

$$E|X(t)| \leq |x| + \int_0^t [|a|E|X(s)| + |b|]ds + E \left| \int_0^t \sigma(X(s))dW(s) \right| + E\check{\xi}(t)$$

$$\begin{aligned}
(2.15) \quad & \leq |x| + |b|t + E\check{\xi}(t+) + \left[E \int_0^t \sigma^2(X(s)) ds \right]^{\frac{1}{2}} + \int_0^t |a|E|X(s)|ds \\
& \leq |x| + |b|t + E\check{\xi}(t+) + \left[1 + E \int_0^t \sigma^2(X(s)) ds \right] + \int_0^t |a|E|X(s)|ds \\
& \leq [|x| + 1 + (|b| + 2K^2)t] + E\check{\xi}(t+) + 2K^2 \int_0^t E|X(s)|^2 ds \\
& \quad + |a| \int_0^t E|X(s)|ds.
\end{aligned}$$

Let $p(t) = [1 + |x| + (|b| + 2K^2)t] + E\check{\xi}(t+) + 2K^2 \int_0^t E|X(s)|^2 ds$. We claim that $p(t) = o(e^{\alpha t})$, as $t \rightarrow \infty$. Indeed, by assumption, $E\check{\xi}(t+) = o(e^{\alpha t})$, so we must only show that

$$E \int_0^t |X(s)|^2 ds = o(e^{\alpha t}), \quad \text{as } t \rightarrow \infty.$$

Define $\phi(t) = e^{-\alpha t} E|X(t)|^2$; then the assumption implies that $\int_0^\infty \phi(t) dt < \infty$. Therefore, a simple application of dominated convergence theorem leads to

$$e^{-\alpha t} E \int_0^t |X(s)|^2 ds = \int_0^\infty 1_{[0,t]}(s) e^{-\alpha(t-s)} \phi(s) ds \longrightarrow 0, \quad \text{as } t \rightarrow \infty.$$

This proves the claim.

Now applying Gronwall's inequality (e.g., cf. [7, eq. (2.7.1)]) to (2.15), we obtain that

$$(2.16) \quad E|X(t)| \leq p(t) + \int_0^t e^{|a|(t-s)} p(s) ds.$$

Note that $p(t) = o(e^{\alpha t})$ and $\alpha > |a|$, so, given $\epsilon > 0$, we can choose $T > 0$ so that $e^{-\alpha s} p(s) < \epsilon$ for $s \geq T$; hence, for some $k_1 > 0$,

$$\begin{aligned}
\int_0^t e^{|a|(t-s)} p(s) ds & \leq \int_0^T e^{|a|(t-s)} p(s) ds + \epsilon \int_T^t e^{|a|(t-s)} \cdot e^{\alpha s} ds \\
& \leq e^{|a|t} k_1 T + e^{|a|t} \cdot \epsilon \cdot \int_T^t e^{(\alpha - |a|)s} ds \\
& \leq e^{|a|t} k_1 T + e^{|a|t} \cdot \epsilon \cdot \frac{e^{(\alpha - |a|)t}}{\alpha - |a|}.
\end{aligned}$$

Since ϵ is arbitrary, we obtain that $\int_0^t e^{|a|(t-s)} p(s) ds = o(e^{\alpha t})$; the consequence then follows from (2.16). \square

Proof of the theorem. Our approach is typical. Let $\xi \in \mathcal{B}$ and write $\xi = \xi^+ - \xi^-$. Let $X(\cdot) = X_\xi^x(\cdot)$ be the corresponding solution of (2.1). Denote the right-continuous version of ξ by $\{\xi(t+), t \geq 0\}$. (The right-continuous version of an adapted left-continuous process $\eta(\cdot)$ is a process $\zeta(\cdot)$ such that, for each $\omega \in \Omega$, $\zeta(t, \omega) = \eta(t+, \omega)$, for all $t \geq 0$ and $\zeta(0-, \omega) = \eta(0, \omega)$. The right-continuity of the filtration $\{\mathcal{F}_t\}$ guarantees that ζ is also adapted.) Define $F(t, x) = e^{-\alpha t} V(x)$, for $(t, x) \in [0, \infty) \times \mathbf{R}$. By the generalized Itô formula (Meyer [11]), we have that

$$e^{-\alpha t} V(X(t+)) = V(X(0+))$$

$$\begin{aligned}
(2.17) \quad & + \int_0^t e^{-\alpha s} [-\alpha V(X(s)) + (aX(s) + b)V'(X(s)) + \frac{1}{2}\sigma^2(X(s))V''(X(s))]ds \\
& + \int_{(0,t]} e^{-\alpha s} V'(X(s))d\xi(s+) + \int_0^t e^{-\alpha s} V'(X(s))\sigma(X(s))dW(s) \\
& + \sum_{0 < s \leq t} [V(X(s+)) - V(X(s)) - V'(X(s))(X(s+) - X(s))].
\end{aligned}$$

By (2.13) we see that the second term on the above right-hand side is no less than $-\int_0^t e^{-\alpha s} h(X(s))ds$. The convexity of V implies that

$$\sum_{0 < s \leq t} [V(X(s+)) - V(X(s)) - V'(X(s))(X(s+) - X(s))] \geq 0, \quad \text{a.s.}$$

So (2.17) becomes

$$\begin{aligned}
(2.18) \quad e^{-\alpha t} V(X(t+)) & \geq V(X(0+)) - \int_0^t e^{-\alpha s} h(X(s))ds \\
& + \int_{(0,t]} e^{-\alpha s} V'(X(s))d\xi(s+) \\
& + \int_0^t e^{-\alpha s} V'(X(s))\sigma(X(s))dW(s).
\end{aligned}$$

Note that the convexity of V also implies that

$$\begin{aligned}
0 & \leq V(X(0+)) - V(X(0)) - V'(X(0))(X(0+) - X(0)) \\
& = V(X(0+)) - V(X(0)) - V'(X(0))(\xi(0+) - \xi(0)) \\
& = V(X(0+)) - V(X(0)) - \int_{\{0\}} V'(X(s))d\xi(s+).
\end{aligned}$$

Therefore

$$\begin{aligned}
(2.19) \quad e^{-\alpha t} V(X(t+)) & \geq V(X(0)) - \int_0^t e^{-\alpha s} h(X(s))ds \\
& + \int_{[0,t]} e^{-\alpha s} V'(X(s))d\xi(s+) \\
& + \int_0^t e^{-\alpha s} V'(X(s))\sigma(X(s))dW(s).
\end{aligned}$$

Define

$$(2.20) \quad M(t) = e^{-\alpha t} V(X(t+)) + \int_{[0,t]} e^{-\alpha s} [d\check{\xi}(s+) + h(X(s))ds],$$

$$(2.21) \quad m(t) = \int_0^t e^{-\alpha s} V'(X(s))\sigma(X(s))dW(s).$$

Some computation from (2.19) yields that

$$\begin{aligned}
(2.22) \quad EM(t) & \geq V(x) + E \int_{[0,t]} e^{-\alpha s} [1 + V'(X(s))]d\xi^+(s+) \\
& + E \int_{[0,t]} e^{-\alpha s} [1 - V'(X(s))]d\xi^-(s+) + Em(t).
\end{aligned}$$

Since $|V'(x)| \leq 1$, (2.22) gives $EM(t) \geq V(x) + Em(t)$.

Observe that, by the definition of $\tilde{\xi}(t+)$, $\int_{[0,t]} e^{-\alpha s} d\tilde{\xi}(s+) = \int_{[0,t]} e^{-\alpha s} d\tilde{\xi}(s)$ for all $t \geq 0$, since the integrand $e^{-\alpha t}$ is continuous. So the expectation of the second term on the right-hand side of (2.20) converges to $V_\xi(x)$ as $t \rightarrow \infty$. Therefore, to finish the proof, we must only show that $\lim_{t \rightarrow \infty} EM(t) = \lim_{t \rightarrow \infty} e^{-\alpha t} EV(X(t+)) + V_\xi(x) = V_\xi(x)$ and $Em(t) = 0$, whenever $V_\xi(x) < \infty$. (If $V_\xi(x) = \infty$, there is nothing to prove.) It is readily seen, however, that $V_\xi(x) < \infty$ implies that $E\tilde{\xi}(t+) = o(e^{\alpha t})$ and $E \int_0^\infty e^{-\alpha t} |X(t)|^2 dt < \infty$; i.e., the assumptions of Lemma 2.2 are satisfied. The latter, together with (2.13) and (2.7), implies that $m(t)$ is a L^2 -martingale, so $Em(t) = 0$ for each $t \geq 0$. On the other hand, by Lemma 2.2, we have that $E|X(t+)| = o(e^{\alpha t})$. It follows immediately that $EV(X(t+)) = o(e^{\alpha t})$, since $V(\cdot)$ is at most of linear growth by (2.13). This leads to the conclusion that $\lim_{t \rightarrow \infty} e^{-\alpha t} EV(X(t+)) = 0$. Therefore $V_\xi(x) \geq V(x)$, $x \in \mathbf{R}$. The remainder of the theorem is obvious, so we are done. \square

Finally, we give a local version of Theorem 2.1, which will be very useful in this paper. Since the proof is virtually identical to that of Theorem 2.1, we omit it.

THEOREM 2.3. *Let V be a C^2 -function defined on \mathbf{R} satisfying (2.14). Let $-\infty < L < B < \infty$ and suppose that V satisfies (2.12), (2.13) on $[L, B]$. Then under assumptions (A1)–(A3), we have that*

$$V(x) \leq \inf_{\xi \in \mathcal{B}_{[L, B]}} V_\xi(x), \quad x \in [L, B].$$

Furthermore, if there exists a $\xi^* \in \mathcal{B}_{[L, B]}$ such that $V(x) = V_{\xi^*}(x)$ for all $x \in [L, B]$, then

$$V(x) = V_{\xi^*}(x) = \inf_{\xi \in \mathcal{B}_{[L, B]}} V_\xi(x), \quad x \in [L, B].$$

3. Some basic results for the ODE related to H-J-B equation. In this section, we study the following ODE related to the H-J-B equation (2.11) under assumptions (A1)–(A3):

$$(3.1) \quad \alpha V(x) = (ax + b)V'(x) + \frac{1}{2}\sigma^2(x)V''(x) + h(x), \quad x \in \mathbf{R}$$

and give some results that serve as lemmas for the main theorem.

We consider the following *free boundary problem*. Find a pair of real numbers $-\infty < L < B < \infty$ and a solution V of (3.1) that is *convex* on $[L, B]$, satisfying the boundary conditions

$$(3.2) \quad V'(L) = -1, \quad V'(B) = 1;$$

$$(3.3) \quad V''(L) = V''(B) = 0.$$

Remark 3.1. For the boundary conditions (3.2) and (3.3), it should be understood first that all the derivatives there are one-sided in the appropriate direction. Then observe that once a solution exists on $[L, B]$, it can actually be extended to be defined on the whole real line by our assumptions on the data. Hence, in the following, the derivatives at the boundary will be the usual two-sided derivatives.

The other observation is that, since σ, h are of class C^2 and since σ is nonvanishing, we can easily check by directly differentiating (3.1) that any solution of (3.1) will be of class C^4 (on the whole real line).

We claim that, under our basic assumptions, the solution to (3.1), (3.2) exists and is unique for any given $L < B$. Indeed, let f, g be two independent solutions to the homogeneous equation

$$(3.4) \quad \frac{1}{2}\sigma^2(x)V''(x) + (ax + b)V'(x) - \alpha V(x) = 0$$

with the boundary conditions

$$(3.5) \quad \begin{aligned} f(0) &= 1; & g(0) &= 0; \\ f'(0) &= 0; & g'(0) &= 1; \end{aligned}$$

then a general solution of (3.1) can be written as

$$(3.6) \quad V(x) = C_1 f(x) + C_2 g(x) - 2 \int_0^x \varphi(x, s) \frac{h(s)}{\sigma^2(s)} ds,$$

where $\varphi(\cdot, s)$ is the solution of (3.4) for $x \geq s$, satisfying

$$(3.7) \quad \varphi(s, s) = 0; \quad \varphi_x(s, s) = 1$$

(cf. [3]). Clearly, the existence and uniqueness of the solution to the boundary problem (3.1), (3.2) for given $L < B$ is equivalent to the fact that

$$(3.8) \quad \begin{vmatrix} f'(L) & g'(L) \\ f'(B) & g'(B) \end{vmatrix} = f'(L)g'(B) - g'(L)f'(B) \neq 0.$$

Let $\Psi(x) = f'(L)g(x) - g'(L)f(x)$; then Ψ is a solution to (3.4) with $\Psi'(L) = 0$. So it follows from the following lemma quoted from Shreve [12] that $\Psi'(B) \neq 0$, i.e., (3.8) holds. (We outline the proof of this lemma in the Appendix for the benefit of the reader.)

LEMMA 3.1. *Suppose that $\alpha > |a|$ and let V be a nonconstant solution to (3.4) defined on some interval $[L, B]$; then*

(a) *If V has a zero in $[L, B]$, then V' has no zero in $[L, B]$;*

(b) *If $V'(\bar{x}) = 0$ for some $\bar{x} \in [L, B]$, then $(x - \bar{x})V(x)V'(x) > 0$, for all $x \in [L, B]$ such that $x \neq \bar{x}$.*

We can now write the explicit formula for C_1, C_2 to solve the boundary problem (3.1), (3.2) for given $L < B$, as follows:

$$(3.9) \quad C_1 = \frac{1}{\Delta} \det \begin{bmatrix} 2I_1(L) - 1 & g'(L) \\ 2I_1(B) + 1 & g'(B) \end{bmatrix}, \quad C_2 = \frac{1}{\Delta} \det \begin{bmatrix} f'(L) & 2I_1(L) - 1 \\ f'(B) & 2I_1(B) + 1 \end{bmatrix},$$

where $I_1(x) = \int_0^x \varphi_x(x, s)(h(s)/\sigma^2(s))ds$, and

$$\Delta = \det \begin{bmatrix} f'(L) & g'(L) \\ f'(B) & g'(B) \end{bmatrix}.$$

We will henceforth denote, for given $L < B$, the solution to (3.1), (3.2) by $V_{L,B}$. (Recall from Remark 3.1 that it is actually defined on \mathbf{R} and is of class C^4 .) The following lemmas give the crucial properties of such solutions.

LEMMA 3.2. *Let $V_{L,B}$ be the solution to (3.1), (3.2) on some interval $[L, B] \subset \mathbf{R}$; then the following statements are equivalent:*

- (1) $V_{L,B}$ is convex on $[L, B]$;
- (2) $|V'_{L,B}(x)| \leq 1$ for all $x \in [L, B]$;
- (3) $V''_{L,B}(L) \geq 0, V''_{L,B}(B) \geq 0$.

Proof. We denote $V = V_{L,B}$.

(1) \implies (2). If V is convex, then V' is increasing, so the boundary condition (3.2) gives $|V'(x)| \leq 1$ for all $x \in [L, B]$.

(2) \implies (3). This is obvious by (3.2).

(3) \implies (1). We prove that $V''(x) \geq 0$ for all $x \in [L, B]$. Differentiating both sides of (3.1) twice and letting $W = V''$, we have that

$$0 = \frac{1}{2}\sigma^2(x)W''(x) + [(ax+b) + (\sigma^2(x))']W'(x) + (2a + \frac{1}{2}(\sigma^2(x))'' - \alpha)W(x) + h''(x).$$

Let $c(x) = 2a + \frac{1}{2}(\sigma^2(x))'' - \alpha$ and

$$(LW)(x) = \frac{1}{2}\sigma^2(x)W''(x) + [(ax+b) + (\sigma^2(x))']W'(x) + c(x)W(x);$$

then we have that $LW = -h'' < 0$ by (A2) and $c < 0$ by (A3). Therefore, by the *maximum principle* (cf. [4]), $W = V''$ has no negative minimum on $[L, B]$. Thus $V''(x) \geq 0$ on $[L, B]$, since otherwise V'' must have a negative minimum by the assumption. The proof is now completed. \square

Note that assumption (A2) implies that there exists a unique pair of real numbers $-\infty < r_1 < 0 < r_2 < \infty$ with $|h'(r_1)| = |h'(r_2)| = \alpha - a$ such that $|h'(x)| < \alpha - a$ for all $x \in (r_1, r_2)$ (see also Remark 2.1). We have the following lemma.

LEMMA 3.3. *Suppose that $[L, B] \subseteq \mathbf{R}$ and $V_{L,B}$ are the same as those in Lemma 3.2 and suppose that $V_{L,B}$ is convex on $[L, B]$; then*

- (i) $V''_{L,B}(B) = 0 \implies h'(B) \geq \alpha - a > 0$ and $B \geq r_2 > 0$;
- (ii) $V''_{L,B}(L) = 0 \implies h'(L) \leq -(\alpha - a) < 0$ and $L \leq r_1 < 0$.

Proof. We only prove (i) (the proof of (ii) is similar). Again, denote $V = V_{L,B}$; as already observed, $V'''(x)$ exists for all x and satisfies

$$\frac{1}{2}\sigma^2(x)V'''(x) + [(ax+b) + \sigma(x)\sigma'(x)]V''(x) + (a-\alpha)V'(x) + h'(x) = 0.$$

Now, letting $x \nearrow B$, we get that

$$(3.10) \quad \frac{1}{2}\sigma^2(B)V'''(B) + (a-\alpha) + h'(B) = 0.$$

Since $V''(x) \geq 0$ for all $x \in [L, B]$ and $V''(B) = 0$, we get that $V'''(B) \leq 0$. Hence the result follows from (3.10), condition (2.9), and the definition of r_2 . \square

We now give a lemma concerning the continuous dependence of the solution on the boundary data. It is easily seen that $C_1 = C_1(L, B), C_2 = C_2(L, B)$ given by (3.9) are continuous functions of L and B for $L < B$, since $\Delta \neq 0$ for all $L < B$. However, it is not clear that, if $V_{L,B}$ is convex on some $[L, B]$, then $V_{L',B'}$ should also be convex on $[L', B']$ for those L' close to L and B' close to B . We have the following lemma.

LEMMA 3.4. *Suppose that, for some $-\infty < L < B < \infty$, the function $V_{L,B}$ is convex on $[L, B]$. If $V_{L,B}''(B) > 0$ (respectively, $V_{L,B}''(L) > 0$), then, for any $\epsilon > 0$, there exist $L', B' \in \mathbf{R}$ with $B < B'; L \leq L' < L + \epsilon < B$ (respectively, $L' < L; L < B - \epsilon < B' \leq B$), such that $V_{L',B'}$ is convex on $[L', B']$.*

Proof. Since $V_{L,B}''(B) > 0$, by the continuity of C_1, C_2 in L, B , we can find that $\delta_0 > 0$ such that the solution $V_{L,B+\delta}$ satisfies $V_{L,B+\delta}''(B+\delta) > 0$ for $0 < \delta < \delta_0$. We may assume that $\delta_0 < \epsilon < 1$.

Now pick $x_0 \in (L, L + \epsilon)$ such that $V'_{L,B}(x_0) > -1$. Such an x_0 always exists; otherwise, by part (2) of Lemma 3.2, $V'_{L,B}(x) \equiv -1$ on $[L, L + \epsilon]$, which implies that $V_{L,B}$ is linear on $[L, L + \epsilon]$. Then, however, $h(x) = \alpha V_{L,B}(x) + (ax + b)$, $x \in [L, L + \epsilon]$ is also linear. This contradicts (2.8).

Let $\epsilon_1 = V'_{L,B}(x_0) + 1 > 0$. Note that the solution family $\{V_{L,B+\delta}, 0 \leq \delta < \delta_0\}$ are all defined on \mathbf{R} , so the explicit form of the solution (3.6), (3.9) and the continuity of $V''_{L,B}$ shows that we can choose $0 < \delta_1 < \delta_0$ such that

$$|V'_{L,B+\delta_1}(x) - V'_{L,B}(x)| < \epsilon_1, \quad x \in [L, B + 1]$$

and

$$V''_{L,B}(x) > 0, \quad x \in [B, B + \delta_1],$$

since $V''_{L,B}(B) > 0$. It follows that $V'_{L,B}(x) \geq 1$ for all $x \in [B, B + \delta_1]$ and therefore $V'_{L,B+\delta_1}(x) > -1$ for all $x \in [x_0, B + \delta_1]$. Let

$$L' = \inf\{u \geq L : V'_{L,B+\delta_1}(x) > -1, u \leq x \leq B + \delta_1\};$$

then $L \leq L' < x_0 < L + \epsilon$.

It is easily seen from the definition of L' that we must have that $V''_{L',B+\delta_1}(L') \geq 0$ and $V'_{L',B+\delta_1}(L') = -1$. By Lemma 3.2, $V_{L',B+\delta_1}$ is convex on $[L', B + \delta_1]$. Therefore, with $B' = B + \delta_1$, the solution $V_{L',B'}$ is just what we want. The case when $V''_{L,B}(L) > 0$ is similar, so we are done. \square

The next question is: When does a convex solution $V_{L,B}$ satisfying (3.1), (3.2) exist? We can prove the following lemma.

LEMMA 3.5. *For any $[L, B] \subset \mathbf{R}$, there exist $L \leq L' < B' \leq B$ such that $V_{L',B'}$ is convex on $[L', B']$.*

Proof. Let $V_{L,B}$ be the solution to (3.1), (3.2) on $[L, B]$. Define

$$\begin{aligned} B' &= \sup\{u : V'_{L,B}(x) \leq 1, \quad L \leq x \leq u\} \wedge B; \\ L' &= \inf\{u : V'_{L,B}(x) \geq -1, \quad u \leq x \leq B'\} \vee L. \end{aligned}$$

By (3.2) and the continuity of $V'_{L,B}$, it is easily seen that $L \leq L' < B' \leq B$; $V'_{L,B}(L') = -1$, $V'_{L,B}(B') = 1$ and $|V'_{L,B}(x)| \leq 1$ for $L' \leq x \leq B'$.

Replacing L by L' and B by B' , we obtain a solution $V_{L',B'}$, which is convex on $[L', B']$ by Lemma 3.2. \square

By Lemma 3.5, we see that, for any $[L, B] \in \mathbf{R}$, the solution $V_{L,B}$ to (3.1), (3.2) has a *convex portion*, which also satisfies (3.2). We concentrate on the totality of such convex portions. Define, for each $x_0 \in \mathbf{R}$,

$$(3.11) \quad \mathcal{A}_{x_0} = \{[L, B] : x_0 \in (L, B) \text{ and there exists a } V_{L,B} \text{ convex on } [L, B]\}.$$

Apparently, for r_1, r_2 defined as those in Lemma 3.3 (and the argument preceding it), there exists $r_1 < x_0 < r_2$ such that $\mathcal{A}_{x_0} \neq \emptyset$. Denote $\mathcal{A} = \mathcal{A}_{x_0}$. (As we see in §4, we may actually take $x_0 = 0$.) We find a unique $[L^*, B^*] \in \mathcal{A}$ such that the corresponding V_{L^*,B^*} satisfies (3.3).

The following lemma is a basic property of \mathcal{A} . We endow a partial order “ \prec ” on \mathcal{A} by usual inclusion; i.e.,

$$[L, B] \prec [L', B'] \iff [L, B] \subseteq [L', B'].$$

LEMMA 3.6. (a) \mathcal{A} is “closed” in the following sense: if $\{[L_n, B_n]\} \subseteq \mathcal{A}$ such that $L_n \rightarrow L; B_n \rightarrow B$, and $-\infty < L < x_0 < B < \infty$, then $[L, B] \in \mathcal{A}$.

(b) Every totally ordered subset of \mathcal{A} has an upper bound.

Proof. (a) Let $\{[L_n, B_n]\} \subseteq \mathcal{A}$ such that $L_n \rightarrow L; B_n \rightarrow B$ for some $-\infty < L < x_0 < B < \infty$. Let C_1^n, C_2^n be the constants in (3.6) with respect to V_{L_n, B_n} determined by (3.11) (with corresponding Δ^n); then it is easily seen that there exist some C_1, C_2 , and Δ such that $C_i^n \rightarrow C_i, i = 1, 2$, and, $\Delta^n \rightarrow \Delta$, since $L < B$. It can then be checked that C_1, C_2 determine a solution $V_{L, B}$ to (3.1), (3.2) on $[L, B]$ via (3.6) such that $V_{L, B}''(L) \geq 0$ and $V_{L, B}''(B) \geq 0$, since $V_{L_n, B_n}''(L_n) \geq 0$ and $V_{L_n, B_n}''(B_n) \geq 0$ for every n . Therefore $V_{L, B}$ is convex on $[L, B]$ by Lemma 3.2; i.e. $[L, B] \in \mathcal{A}$.

(b) Let $\{[L_\lambda, B_\lambda] : \lambda \in \Lambda\}$ be a totally ordered subset of \mathcal{A} ; then there exist $\mathbf{L} < \mathbf{B}$ such that $(\mathbf{L}, \mathbf{B}) = \cup_\lambda (L_\lambda, B_\lambda)$.

It can be proved that $-\infty < \mathbf{L} < \mathbf{B} < \infty$ (we defer the proof to next section, Lemma 4.2). Moreover, since, for each λ , $L_\lambda < x_0 < B_\lambda$, then $\mathbf{L} < x_0 < \mathbf{B}$.

If Λ is a finite set or $\{[L_\lambda, B_\lambda]\}$ has a maximum element, then there is nothing to prove. So assume that Λ is infinite and that there is no maximum element in the family. Then we can find a sequence

$$[L_1, B_1] \subseteq [L_2, B_2] \subseteq \cdots$$

such that $L_n \searrow \mathbf{L}, B_n \nearrow \mathbf{B}$. By part (a), $[\mathbf{L}, \mathbf{B}] \in \mathcal{A}$. It is clear that $[\mathbf{L}, \mathbf{B}]$ is the upper bound of the family $\{[L_\lambda, B_\lambda]\}$. \square

Now, by Lemma 3.6 and Zorn's lemma, we see that \mathcal{A} has a maximal element. We should note that the maximal element is not unique, since \mathcal{A} is only a partially ordered set. However, we may now define a subset of \mathcal{A} as follows:

$$(3.12) \quad \mathcal{A}_{\max} = \{ \text{all maximal elements in } \mathcal{A} \}.$$

The previous argument shows that $\mathcal{A}_{\max} \neq \emptyset$. We are mostly interested in this set later.

To end this section, we present a simple but important property of \mathcal{A}_{\max} .

LEMMA 3.7. For any $[L, B] \in \mathcal{A}_{\max}$, we have that

$$(3.13) \quad V_{L, B}''(L) \cdot V_{L, B}''(B) = 0.$$

Proof. First, note that, for any $[L, B] \in \mathcal{A}$, we have that $V_{L, B}''(L) \geq 0; V_{L, B}''(B) \geq 0$. So, if the conclusion is not true, then we can find an $[L, B] \in \mathcal{A}_{\max} \subseteq \mathcal{A}$ such that

$$V_{L, B}''(L) > 0; \quad V_{L, B}''(B) > 0.$$

Then, by the continuous dependence of the solution on L, B , we can find an $\epsilon > 0$, so that $V_{L-\epsilon, B+\epsilon}$ exists on $[L-\epsilon, B+\epsilon]$ and satisfies

$$V_{L-\epsilon, B+\epsilon}''(L-\epsilon) \geq 0; \quad V_{L-\epsilon, B+\epsilon}''(B+\epsilon) \geq 0.$$

So Lemma 3.2 implies that $V_{L-\epsilon, B+\epsilon}$ is convex, and then $[L-\epsilon, B+\epsilon] \in \mathcal{A}$, since $x_0 \in (L, B) \subset (L-\epsilon, B+\epsilon)$, but this contradicts the maximality of $[L, B]$. \square

4. Main theorems. In this section, we give our main results. The first theorem is relatively simple, but we still prove it for completeness. The remainder of the section is devoted to the second theorem, which is more involved. We prove that the

principle of smooth fit always holds under our setting, and then the first theorem leads to the existence of the optimal control. Finally, we give a brief description of the optimal reflecting barriers.

First, let $[L, B] \subset \mathbf{R}$ and let $X^x(\cdot)$ denote the diffusion process starting at $x \in [L, B]$, satisfying

$$dX^x(t) = (aX^x(s) + b)ds + \sigma(X^x(s))dW(s),$$

with reflection at L and B . Then, following §23 in [5], we have two adapted, continuous, nondecreasing processes $\xi_L(\cdot)$ and $\xi_B(\cdot)$, which are zero at $t = 0$, such that, for all $t \geq 0$,

$$(4.1) \quad X^x(t) = x + \int_0^t (aX^x(s) + b)ds + \int_0^t \sigma(X^x(s))dW(s) + \xi_L(t) - \xi_B(t)$$

and

$$\xi_L(t) = \int_0^t 1_{\{X(s)=L\}} d\xi_L(s), \quad \xi_B(t) = \int_0^t 1_{\{X(s)=B\}} d\xi_B(s).$$

Denote such solution by $X_\xi^x(\cdot)$. Let f be a solution to (3.1) on $[L, B]$ and let $F(t, x) = e^{-\alpha t} f(x)$. Applying Itô's formula to the function F , we obtain that

$$(4.2) \quad \begin{aligned} f(x) = & f'(B)E \int_0^\infty e^{-\alpha t} d\xi_B(t) - f'(L)E \int_0^\infty e^{-\alpha t} d\xi_L(t) \\ & + E \int_0^\infty e^{-\alpha t} h(X_{\xi_{L,B}}^x(t))dt, \end{aligned}$$

where $\xi_{L,B} = \xi_L - \xi_B$. (See also [13, Lemma 2.1]. Note that it was also proved there that both $E \int_0^\infty e^{-\alpha t} d\xi_B(t)$ and $E \int_0^\infty e^{-\alpha t} d\xi_L(t)$ are finite.) If $V_{L,B}$ is a solution to (3.1), (3.2), then, with $f = V_{L,B}$, (4.2) becomes

$$(4.3) \quad V_{L,B}(x) = E \int_0^\infty e^{-\alpha t} [d\check{\xi}_{L,B}(t) + h(X_{\xi_{L,B}}^x(t))dt].$$

Namely, $\xi_{L,B}$ yields the cost function $V_{L,B}$. We now state our main theorems.

THEOREM 4.1. *Suppose that there exists $[L^*, B^*] \subset \mathbf{R}$ and a solution V_{L^*, B^*} to (3.1)–(3.3) on $[L^*, B^*]$; then*

$$(4.4) \quad V^*(x) = \begin{cases} (L^* - x) + V_{L^*, B^*}(L^*), & x < L^*; \\ V_{L^*, B^*}(x), & L^* \leq x \leq B^*; \\ (x - B^*) + V_{L^*, B^*}(B^*), & x > B^* \end{cases}$$

is the value function, and the optimal control is given by $\xi_x^ = \{\xi_x^*(t) : t \geq 0\}$ satisfying $\xi_x^*(0) = 0$, and, for $t > 0$,*

$$(4.5) \quad \xi_x^*(t) = \begin{cases} (L^* - x) + \xi_{L^*}^*(t) - \xi_{B^*}^*(t), & x < L^*; \\ \xi_{L^*}^*(t) - \xi_{B^*}^*(t), & L^* \leq x \leq B^*; \\ (B^* - x) + \xi_{L^*}^*(t) - \xi_{B^*}^*(t), & x > B^*. \end{cases}$$

It is obvious that ξ_x^* is left-continuous, and, for $x < L^*$ ($x > B^*$), ξ_x^* has an initial jump, which makes the process $X^* = X_{\xi_x^*}^x(\cdot)$ jump to L^* (B^*) and then proceeds as a reflected diffusion on $[L^*, B^*]$. This is just the usual idea used by many authors (cf. [1], [6], [8], [13]). Observe also that Theorem 4.1 depends heavily on the existence of the interval $[L^*, B^*]$ and the corresponding convex solution V_{L^*, B^*} . In some cases, the nonexistence of such an interval leads to the nonexistence of the optimal policy (cf. Shreve, Lehoczky, and Gaver [13]). However, the next theorem gives an affirmative answer to the question of the existence of such interval in our setting as well as the existence of the convex C^2 -solution to the variational inequality (2.11).

THEOREM 4.2. *Let assumptions (A1)–(A3) hold. Then there exists a unique interval $[L^*, B^*] \subset \mathbf{R}$ on which there exists a unique, convex solution of (3.1)–(3.3). Furthermore, the variational inequality (2.11) admits a unique convex C^2 -solution, which gives the value function of the control problem (2.1)–(2.3).*

Remark. By setting $a = 0$, $\sigma(\cdot) \equiv \sigma(\text{constant})$, we see that our result contains the corresponding one in [8] as a special case.

Proof of Theorem 4.1. It is readily seen that the control ξ_x^* yields the cost function V^* defined by (4.4), so we need only show that V^* is the optimal cost.

Since $\xi^* \in \mathcal{B}$, we have that

$$(4.6) \quad V^*(x) \geq \inf_{\xi \in \mathcal{B}} V_{\xi}(x), \quad x \in \mathbf{R}.$$

On the other hand, by the assumption of the theorem, we see that $V^* \in C^2(\mathbf{R})$ and

$$(4.7) \quad V^{*'}(x) = \begin{cases} -1, & x < L^*; \\ V'_{L^*, B^*}(x), & L^* \leq x \leq B^*; \\ 1, & x > B^*; \end{cases}$$

$$(4.8) \quad V^{*''}(x) = \begin{cases} 0, & x < L^* \text{ or } x > B^*; \\ V''_{L^*, B^*}(x), & L^* \leq x \leq B^*. \end{cases}$$

By Lemma 3.2, (2.12) and (2.13) are satisfied. We now verify (2.14). If $x \in [L^*, B^*]$, there is nothing to prove. Let $x > B^*$; then, by the definition of V^* , we have that

$$(4.9) \quad \alpha V^*(x) = \alpha(x - B^*) + \alpha V_{L^*, B^*}(B^*).$$

Since V_{L^*, B^*} is a solution of (3.1)–(3.3), we have, at $x = B^*$, that

$$(4.10) \quad \alpha V_{L^*, B^*}(B^*) = \frac{1}{2}\sigma^2(B^*)V^{*''}(B^*) + (aB^* + b)V^{*'}(B^*) + h(B^*) = (aB^* + b) + h(B^*).$$

Thus (4.9) becomes $\alpha V^*(x) = \alpha(x - B^*) + (aB^* + b) + h(B^*)$. Therefore a simple computation shows that

$$(4.11) \quad \alpha V^*(x) \leq (ax + b)V^{*'}(x) + \frac{1}{2}\sigma^2(x)V^{*''}(x) + h(x), \quad x > B^*$$

is equivalent to

$$(4.12) \quad (\alpha - a) \leq \frac{h(x) - h(B^*)}{x - B^*}, \quad x > B^*.$$

Since h is strictly convex, h' is increasing. Therefore, for any $x > B^*$,

$$\frac{h(x) - h(B^*)}{x - B^*} = h'(\theta) > h'(B^*) \geq \alpha - a,$$

where $\theta \in (B^*, x)$ and the last inequality is due to Lemma 3.3 (i). Thus we have proved (2.14) for $x > B^*$. The case when $x < L^*$ is similar, so (2.14) is verified. By Theorem 2.1, we have that

$$V^*(x) \leq \inf_{\xi \in \mathcal{B}} V_\xi(x), \quad x \in \mathbf{R}.$$

The proof is now complete. \square

For the proof of Theorem 4.2, we first prove some lemmas. Our purpose here is to find an interval $[L^*, B^*] \subseteq \mathbf{R}$ satisfying the conditions of Theorem 4.1. The candidate is chosen from the set \mathcal{A}_{\max} defined by (3.12). We now take a closer look at the sets $\mathcal{A}, \mathcal{A}_{\max}$ defined by (3.11), (3.12).

Define

$$\begin{aligned} \mu &= \sup\{B : \exists [L, B] \in \mathcal{A}\}; \\ \nu &= \inf\{L : \exists [L, B] \in \mathcal{A}\}. \end{aligned}$$

Then we have the following lemma.

LEMMA 4.3. *It holds that*

$$(4.13) \quad -\infty < \nu < x_0 < \mu < \infty,$$

where x_0 is such that $x_0 \in (r_1, r_2)$ and $\mathcal{A} = \mathcal{A}_{x_0} \neq \emptyset$.

Proof. $\nu < x_0 < \mu$ is obvious by the definition of \mathcal{A} ; the proof of the first inequality is the same as that of the last one, so we only prove $\mu < \infty$.

Suppose not; then there exists a sequence $\{[L_n, B_n]\}_{n=0}^\infty \subseteq \mathcal{A}$ such that $B_n \nearrow \infty$. Therefore we can choose a $\delta > 0$ such that $[x_0, x_0 + \delta] \subseteq [L_n, B_n]$ for all $n > 0$. Since, for each n , V_{L_n, B_n} is the convex solution to (3.1), (3.2) on $[L_n, B_n]$, by Theorem 2.3, with the argument in the beginning of this section and part (1) of Remark 2.1, we have that

$$0 \leq V_{L_n, B_n}(x) = \inf_{\xi \in \mathcal{B}_{[L_n, B_n]}} V_\xi(x), \quad n \geq 0, \quad x \in [L_n, B_n].$$

Since $\mathcal{B}_{[x_0, x_0 + \delta]} \subset \mathcal{B}_{[L_n, B_n]}$, for all n , we also have that

$$\inf_{\xi \in \mathcal{B}_{[L_n, B_n]}} V_\xi(x) \leq \inf_{\xi \in \mathcal{B}_{[x_0, x_0 + \delta]}} V_\xi(x) \quad n \geq 0; \quad x \in [x_0, x_0 + \delta].$$

Therefore

$$0 \leq V_{L_n, B_n}(x) \leq \inf_{\xi \in \mathcal{B}_{[x_0, x_0 + \delta]}} V_\xi(x), \quad x \in [x_0, x_0 + \delta].$$

In particular, we have that $0 \leq V_{L_n, B_n}(x_0) \leq v = \inf_{\xi \in \mathcal{B}_{[x_0, x_0 + \delta]}} V_\xi(x_0)$. Since $|V'_{L_n, B_n}(x)| \leq 1$ for $x \in [L_n, B_n]$, we get that

$$(4.14) \quad |V_{L_n, B_n}(x)| \leq v + |x - x_0|, \quad x \in [L_n, B_n].$$

However, the convexity of V_{L_n, B_n} (i.e., $V''_{L_n, B_n} \geq 0$) on $[L_n, B_n]$ gives

$$\begin{aligned}\alpha V_{L_n, B_n}(x) &= (ax + b)V'_{L_n, B_n}(x) + \frac{1}{2}\sigma^2 V''_{L_n, B_n}(x) + h(x) \\ &\geq (ax + b)V'_{L_n, B_n}(x) + h(x), \quad x \in [L_n, B_n].\end{aligned}$$

Hence, by (4.13) and the fact that $|V'_{L_n, B_n}(x)| \leq 1$ for $x \in [L_n, B_n]$, for some $K > 0$,

$$0 \leq h(x) \leq K(1 + |x|), \quad x \in [L_n, B_n].$$

This is impossible, since $B_n \nearrow \infty$ and h is of at least quadratic growth by (2.8). The contradiction shows that $\mu < \infty$. Similarly, we have that $\nu > -\infty$. This completes the proof. \square

Now define

$$(4.15) \quad \mathcal{A}_1 = \{[L, B] \in \mathcal{A}_{\max} : V''_{L, B}(B) = 0\};$$

$$(4.16) \quad \mathcal{A}_2 = \{[L, B] \in \mathcal{A}_{\max} : V''_{L, B}(L) = 0\}.$$

By Lemma 3.7, $\mathcal{A}_{\max} = \mathcal{A}_1 \cup \mathcal{A}_2$.

LEMMA 4.4. *It holds that $\mathcal{A}_1 \neq \emptyset$, $\mathcal{A}_2 \neq \emptyset$.*

Proof. Since $\mathcal{A}_{\max} \neq \emptyset$, one of \mathcal{A}_1 or \mathcal{A}_2 must be nonempty. Suppose that $\mathcal{A}_2 \neq \emptyset$, but $\mathcal{A}_1 = \emptyset$; then $\mathcal{A}_{\max} = \mathcal{A}_2$, and, for any $[L, B] \in \mathcal{A}_{\max}$, we must have that $V''_{L, B}(L) = 0$; $V''_{L, B}(B) > 0$. Let

$$b = \sup\{B : \exists [L, B] \in \mathcal{A}_{\max}\};$$

then $b \leq \mu < \infty$, and there exists a sequence $\{[L_n, B_n]\} \subseteq \mathcal{A}_{\max}$ such that $B_n \nearrow b$. Since $\nu \leq L_n < x_0$ for any n , along a subsequence (may assume itself), we have that $L_n \rightarrow l$ for some $l \leq x_0$. Observe that if $V''_{L_n, B_n}(L_n) = 0$ for all n , we must have that $V''_{l, b}(l) = 0$. By part (ii) of Lemma 3.3 and part (a) Lemma 3.6, $V_{l, b} \in \mathcal{A}$.

Now let $[L^*, B^*]$ be a maximum element containing $[l, b]$; then we must have that $B^* = b$ and $V''_{L^*, B^*}(B^*) = 0$ by the definition of b and Lemma 3.4. However, this contradicts $\mathcal{A}_1 = \emptyset$. The similar argument shows that $\mathcal{A}_1 \neq \emptyset$, but $\mathcal{A}_2 = \emptyset$ is also impossible; so the lemma is proved. \square

It is now clear that we may succeed in proving Theorem 4.2 if we can find an element in $\mathcal{A}_1 \cap \mathcal{A}_2$. To this end, we need the following lemma.

LEMMA 4.5. *Suppose that $[L, B], [L', B'] \in \mathcal{A}_{\max}$ such that $L \leq L' < B \leq B'$ and $V''_{L, B}(B) = V''_{L', B'}(L') = 0$; then $L = L'$; $B = B'$. Consequently, $[L, B] = [L', B'] \in \mathcal{A}_1 \cap \mathcal{A}_2$.*

Proof. First, note that either $L < L', B < B'$ or $L = L', B = B'$ must hold, since both intervals are maximal elements. So we only must prove that the first case is impossible.

Suppose that $L < L', B < B'$. We show that this leads to a contradiction. Let

$$V(x) = \begin{cases} V_{L, B}(x), & x \leq B; \\ (x - B) + V_{L, B}(B), & B \leq x. \end{cases}$$

Then we have that $V \in C^2$, since $V''_{L, B}(B) = 0$. By Lemma 3.3 (ii), we have that $h'(B) \geq \alpha - a$, which leads to

$$\alpha V(x) \leq (ax + b)V'(x) + \frac{1}{2}\sigma^2(x)V''(x) + h(x), \quad x \in \mathbf{R}$$

following the argument in the proof of Theorem 4.1. Clearly, $V(x)$ satisfies (2.11), (2.12) on $[L, B']$; therefore, by Theorem 2.3, we obtain that

$$V(x) \leq V_{\xi}(x); \quad \xi \in \mathcal{B}_{[L, B']}, \quad x \in [L, B'].$$

Now, for $x \in [L', B'] \subset [L, B']$, let $\xi = \xi_{L'} - \xi_{B'} \in \mathcal{B}_{[L', B']} \subseteq \mathcal{B}_{[L, B']}$, where $\xi_{L'}, \xi_{B'}$ are defined in the beginning of this section. Then we see that $V_{\xi}(x) = V_{L', B'}(x) = E \int_0^{\infty} e^{-\alpha t} [d\tilde{\xi}(t) + h(X_{\xi}^x(t))dt]$. So we get that $V(x) \leq V_{L', B'}(x)$, for all $x \in [L', B']$. In particular, we have that

$$V_{L, B}(x) \leq V_{L', B'}(x), \quad x \in [L', B'] \cap [L, B].$$

Similarly, replacing V by the function

$$V(x) = \begin{cases} (L' - x) + V_{L', B'}(L'), & x \leq L'; \\ V_{L', B'}(x), & L' \leq x, \end{cases}$$

we can also show that

$$V_{L', B'}(x) \leq V_{L, B}(x), \quad x \in [L', B'] \cap [L, B],$$

which gives $V_{L, B} \equiv V_{L', B'}$ on $[L', B'] \cap [L, B] = [L', B] \supseteq [r_1, r_2]$ by Lemma 3.3. Hence the uniqueness of the solution to the Cauchy problem of the ODE implies that $V_{L, B} \equiv V_{L', B'}$, but this implies that $[L, B'] \in \mathcal{A}$, contradicting the maximality of $[L', B']$ and $[L, B]$. \square

Define

$$(4.17) \quad \mathbf{B}_1 = \inf\{B : \exists [L, B] \in \mathcal{A}_1\};$$

$$(4.18) \quad \mathbf{L}_2 = \sup\{L : \exists [L, B] \in \mathcal{A}_2\}.$$

Then there exists a sequence $\{[L_n, B_n]\} \subseteq \mathcal{A}_{\max}$ such that $B_n \searrow \mathbf{B}_1$. Since every $[L_n, B_n]$ is a maximal element, $\{L_n\}$ must also be decreasing, and is bounded below by μ . Therefore $L_n \searrow L_1$ for some $L_1 < x_0$. By Lemma 3.6 (a), we have that $[L_1, \mathbf{B}_1] \in \mathcal{A}$. Let $[L_1, \mathbf{B}_1]$ be the maximal element in \mathcal{A} containing $[L_1, \mathbf{B}_1]$; we claim that $\mathbf{B}_1 = \mathbf{B}_1$. Indeed, suppose that $\mathbf{B}_1 > \mathbf{B}_1$; then, for n large enough, we should have that $L_n > L_1 \geq \mathbf{L}_1; \mathbf{B}_1 < B_n < \mathbf{B}_1$, namely, $[L_n, B_n]$ is properly contained in $[L_1, \mathbf{B}_1]$. This contradicts their maximalities. Therefore we may now write $[L_1, \mathbf{B}_1]$ as the maximal element containing $[L_1, \mathbf{B}_1]$.

Similarly, we can find a $\mathbf{B}_2 > \mathbf{L}_2$ such that $[\mathbf{L}_2, \mathbf{B}_2]$ is a maximal element. The following lemma is final.

LEMMA 4.6. *It holds that $[\mathbf{L}_1, \mathbf{B}_1] = [\mathbf{L}_2, \mathbf{B}_2]$.*

Proof. Since $[\mathbf{L}_1, \mathbf{B}_1]$ and $[\mathbf{L}_2, \mathbf{B}_2]$ are both maximal, we either have Case 1: $\mathbf{L}_1 < \mathbf{L}_2, \mathbf{B}_1 < \mathbf{B}_2$ or Case 2: $\mathbf{L}_1 > \mathbf{L}_2, \mathbf{B}_1 > \mathbf{B}_2$, if they are not identical.

Suppose that $\mathbf{L}_1 < \mathbf{L}_2$ and $\mathbf{B}_1 < \mathbf{B}_2$; then, by the definition of \mathbf{B}_1 , we can find that $[L, B] \in \mathcal{A}_1 \subseteq \mathcal{A}_{\max}$ such that $\mathbf{B}_1 \leq B < \mathbf{B}_2$. By the maximality of $[\mathbf{L}_1, \mathbf{B}_1]$ and $[\mathbf{L}_2, \mathbf{B}_2]$, we must have that $\mathbf{L}_1 \leq L < \mathbf{L}_2$. Similarly, by the definition of \mathbf{L}_2 , we can now find that $[L', B'] \in \mathcal{A}_2 \subseteq \mathcal{A}_{\max}$ such that $L < L' \leq \mathbf{L}_2$, and then $B < B' \leq \mathbf{B}_2$ because $[L, B]$ is also maximal. Now, by the definition of \mathcal{A}_1 and \mathcal{A}_2 , we have that $V''_{L', B'}(L') = 0, V''_{L, B}(B) = 0$, which contradicts Lemma 4.5. So Case 1 is impossible.

Suppose that $\mathbf{L}_1 > \mathbf{L}_2$ and $\mathbf{B}_1 > \mathbf{B}_2$. Let $V_1 = V_{\mathbf{L}_1, \mathbf{B}_1}; V_2 = V_{\mathbf{L}_2, \mathbf{B}_2}$. By the definitions of \mathbf{B}_1 and \mathbf{L}_2 , we must have that $V_1''(\mathbf{L}_1) > 0; V_2''(\mathbf{B}_2) > 0$. Then,

by Lemma 3.4, however, there exists a $[L, B] \in \mathcal{A}$ such that $\mathbf{L}_2 < L < \mathbf{L}_1; \mathbf{B}_2 < B < \mathbf{B}_1$. Let $[\hat{L}, \hat{B}]$ be the maximal element containing $[L, B]$; then it is easily seen that $\mathbf{L}_1 > \hat{L} > \mathbf{L}_2$ and $\mathbf{B}_1 > \hat{B} > \mathbf{B}_2$ still hold, since $[\mathbf{L}_i, \mathbf{B}_i], i = 1, 2$ and $[\hat{L}, \hat{B}]$ are all maximal elements. Now, however, by the definition of \mathbf{B}_1 and \mathbf{L}_1 , we must have that $V''_{\hat{L}, \hat{B}}(\hat{B}) > 0$, since $\hat{B} < \mathbf{B}_1$, and $V''_{\hat{L}, \hat{B}}(\hat{L}) > 0$ since $\hat{L} > \mathbf{L}_2$. Therefore $V''_{\hat{L}, \hat{B}}(\hat{L}) \cdot V''_{\hat{L}, \hat{B}}(\hat{B}) > 0$, which contradicts Lemma 3.7. So Case 2 is also impossible. Namely, $[\mathbf{L}_1, \mathbf{B}_1]$ and $[\mathbf{L}_2, \mathbf{B}_2]$ must be identical. \square

Proof of Theorem 4.2. Let $L^* = \mathbf{L}_1 = \mathbf{L}_2; B^* = \mathbf{B}_1 = \mathbf{B}_2$ and let $V^*(x) = V_{L^*, B^*}(x)$, $x \in [L^*, B^*]$. We first prove that V^* is the convex solution of (3.1)–(3.3) on $[L^*, B^*]$.

That V^* is the convex solution of (3.1), (3.2) is clear by the definition of $[L^*, B^*]$. So we must only verify (3.3). By Lemma 3.8, we have that $V^{**}(L^*) \cdot V^{**}(B^*) = 0$. We assume that $V^{**}(L^*) = 0$. The convexity of V^* implies that $V^{**}(B^*) \geq 0$. Suppose that $V^{**}(B^*) > 0$. Recall that $B^* = \mathbf{B}_1$; hence there exists a sequence $\{[L_n, B_n]\} \subseteq \mathcal{A}_1$ such that $B_n \searrow B^*$, and there exists an $\bar{L} \in (-\infty, x_0]$ such that $L_n \searrow \bar{L}$. If $\bar{L} > L^*$, then Lemma 3.4 and the maximality of $[L^*, B^*]$ allow the existence of an element $[L, B] \in \mathcal{A}_{\max}$ such that $L^* < L < \bar{L}$ and $B^* < B$, which is impossible because then we can find that $[L_n, B_n] \subset [L, B]$ for n large enough, which contradicts the maximality of $[L_n, B_n]$. Therefore $\bar{L} = L^*$, and so $V^{**}(B^*) = \lim_{n \rightarrow \infty} V''_{L_n, B_n}(B_n) = 0$, a contradiction. Thus the existence of the interval $[L^*, B^*]$ is proved.

To see the uniqueness, let $[L^{**}, B^{**}]$ be another such interval. By Lemma 3.3, we have that $x_0 \in (L^{**}, B^{**})$; so $[L^{**}, B^{**}] \in \mathcal{A}$. The same proof as that of Lemma 4.6 shows that neither $L^* < L^{**}; B^* < B^{**}$ nor $L^* > L^{**}; B^* > B^{**}$ is possible. The maximality of $[L^*, B^*]$ shows that it cannot be contained in $[L^{**}, B^{**}]$; so the only possible case is $[L^{**}, B^{**}] \subseteq [L^*, B^*]$. Since $\mathcal{B}_{[L^{**}, B^{**}]} \subseteq \mathcal{B}_{[L^*, B^*]}$, we have that

$$V_{L^*, B^*}(x) \leq V_{L^{**}, B^{**}}(x), \quad x \in [L^{**}, B^{**}].$$

On the other hand, let

$$V^{**}(x) = \begin{cases} (L^{**} - x) + V_{L^{**}, B^{**}}(L^{**}), & L^* \leq x \leq L^{**}; \\ V_{L^{**}, B^{**}}(x), & L^{**} \leq x \leq B^{**}; \\ (x - B^{**}) + V_{L^{**}, B^{**}}(B^{**}), & B^{**} \leq x \leq L^*. \end{cases}$$

Then Theorem 2.3 shows that

$$V_{L^{**}, B^{**}}(x) \leq V_{L^*, B^*}(x), \quad x \in [L^{**}, B^{**}].$$

It follows immediately from the uniqueness of the ODE that $V^* \equiv V_{L^{**}, B^{**}}$, and so $L^* = L^{**}; B^* = B^{**}$. Thus the first part of the theorem is proved.

To prove the second part, let $[L^*, B^*], V_{L^*, B^*}$ be those in the first part; then (4.4) in Theorem 4.1 presents a solution to the variational inequality (2.11). So we must only prove the uniqueness.

Let $V(\cdot)$ be any convex C^2 -solution to (2.11); then, for any $x \in \mathbf{R}$, either $|V'(x)| = 1$ or $V(x)$ satisfy (3.1). Let $\mathcal{C} = \{x : |V'(x)| < 1\}$; then V must satisfy (3.1) on \mathcal{C} . So the growth condition of h implies that \mathcal{C} is bounded (see also Lemma 4.3). Moreover, the monotonicity of $V'(\cdot)$ shows that $\mathcal{C} = (L, B)$, where

$$L = \inf\{x : x \in \mathcal{C}\}, \quad B = \sup\{x : x \in \mathcal{C}\};$$

so that $V'(L) = -1$, $V'(B) = 1$. Then, however, we must have that $V'(x) \equiv -1$ for $x \leq L$ and $V'(x) \equiv 1$ for $x \geq B$, which leads to $V''(L) = 0$ and $V''(B) = 0$ because V is C^2 . Therefore, by the first part, $[L, B]$ is unique ($= [L^*, B^*]$), and V must be linear outside $[L, B]$; thus V must be of the form (4.4). This proves the uniqueness, and then the theorem. \square

4.1. A discussion of determining the optimal reflecting barriers. Having worked diligently to get the existence and uniqueness of the optimal reflecting barriers L^* and B^* , we now present a somewhat “explicit” way of determining these two points via a system of (maybe transcendental) equations. The scheme that we use is similar to that in [6].

For any given $-\infty < L < B < \infty$, let ϕ_1, ϕ_2 be two independent solutions to (3.4) satisfying the boundary conditions

$$\begin{aligned}\phi'_1(L) &= 1; & \phi'_2(L) &= 0; \\ \phi'_1(B) &= 0; & \phi'_2(B) &= 1,\end{aligned}$$

and let G be a special solution of (3.1) satisfying the boundary condition $G(L) = G(B) = 0$. Such solutions exist by the argument given in the beginning of §3. Set

$$(4.19) \quad \Psi(x) = G(x) - [1 + G'(L)]\phi_1(x) + [1 - G'(B)]\phi_2(x).$$

We can easily check that Ψ is the solution to (3.1) and (3.2). Differentiating (4.19) twice, using the facts that G satisfies (3.1) and ϕ_1, ϕ_2 satisfy (3.4), and setting $\Psi''(L) = \Psi''(B) = 0$, $a(x) = ax + b$, we get that

$$\begin{aligned}0 = \Psi''(L) &= \frac{2}{\sigma^2(L)} \{-h(L) + a(L) + \alpha[(1 - G'(B))\phi_2(L) - (1 + G'(L))\phi_1(L)]\}, \\ 0 = \Psi''(B) &= \frac{2}{\sigma^2(B)} \{-h(B) - a(B) + \alpha[(1 - G'(B))\phi_2(B) - (1 + G'(L))\phi_1(B)]\},\end{aligned}$$

or, equivalently,

$$(4.20) \quad h(L) = a(L) + \alpha[(1 - G'(B))\phi_2(L) - (1 + G'(L))\phi_1(L)],$$

$$(4.21) \quad h(B) = -a(B) + \alpha[(1 - G'(B))\phi_2(B) - (1 + G'(L))\phi_1(B)].$$

Theorem 4.2 shows that (4.20), (4.21) admit a unique solution (L^* and B^*), which gives the optimal reflecting barriers. In the case when $a(x) \equiv a$; $\sigma(x) \equiv \sigma$ are both constants, we can write

$$(4.22) \quad \phi_1(x) = C_1 e^{\lambda_1 x} + C_2 e^{\lambda_2 x}; \quad \phi_2(x) = \hat{C}_1 e^{\lambda_1 x} + \hat{C}_2 e^{\lambda_2 x}$$

with

$$\begin{aligned}\lambda_1 &= \frac{-a + \sqrt{a^2 + 2\alpha\sigma^2}}{\sigma^2}; & \lambda_2 &= \frac{-a - \sqrt{a^2 + 2\alpha\sigma^2}}{\sigma^2}, \\ C_1 &= \frac{\exp(\lambda_2 B)}{\lambda_1 \Delta}; & C_2 &= -\frac{\exp(\lambda_1 B)}{\lambda_2 \Delta}; \\ \hat{C}_1 &= -\frac{\exp(\lambda_2 L)}{\lambda_1 \Delta}; & \hat{C}_2 &= \frac{\exp(\lambda_1 L)}{\lambda_2 \Delta},\end{aligned}$$

where $\Delta = \exp(\lambda_2 B + \lambda_1 L) - \exp(\lambda_1 B + \lambda_2 L)$, and, finally,

$$G(x) = c_1 e^{\lambda_1 x} + c_2 e^{\lambda_2 x} - \frac{2}{\sqrt{a^2 + 2\alpha\sigma^2}} \int_0^x e^{-(a/\sigma^2)(x-s)} \sinh\left(\frac{\sqrt{a^2 + 2\alpha\sigma^2}}{\sigma^2}(x-s)\right) h(s) ds,$$

where c_1, c_2 are chosen so that $G(L) = G(B) = 0$. Of course, we can pose more conditions on a, σ , and h (e.g., $a = 0, \sigma = 1$, or h is symmetric and so on) to make (4.20), (4.21) more explicit. For example, if $a = 0, \sigma = 1$, (4.22) becomes

$$\phi_1(x) = -\frac{1}{\sqrt{2\alpha}} \frac{\cosh \sqrt{2\alpha}(B-x)}{\sinh \sqrt{2\alpha}(B-L)}; \quad \phi_2(x) = \frac{1}{\sqrt{2\alpha}} \frac{\cosh \sqrt{2\alpha}(x-L)}{\sinh \sqrt{2\alpha}(B-L)},$$

and so on. For the simpler case—when h is an even function—Karatzas [8] had a transcendental equation to determine B^* and L^* ($= -B^*$) by a slightly different method. However, by the uniqueness of such solution, (4.20) and (4.21) would also give the same answer.

5. The linear case. In this section, we consider the case when σ is also linear. More precisely, we assume that $a(x) = ax + b; \sigma(x) = \theta(ax + b)$, where a, b, θ are constants and $a \neq 0, \theta \neq 0$.

Clearly, the basic assumption (A1) is partially violated, since now σ possesses a zero at $x = -b/a$. The major disadvantage of this violation is that the ODE related to the H-J-B equation now has a singularity at the zero of σ . We then wonder whether the value function is still C^2 . However, we prove directly that, under the extra condition on the position of the “vertex” of function h (condition (5.4)), such a singularity is removable. Namely, there still exists a convex C^2 -solution to the variational inequality (2.11), which is now of the form

$$(5.1) \quad [\alpha V(x) - \tfrac{1}{2}\theta^2(ax+b)^2 V''(x) - (ax+b)V'(x) - h(x)] \vee [|V'(x)| - 1] = 0, \quad x \in \mathbf{R}.$$

We will also derive the explicit formula for determining the smooth fitting points (it might be a transcendental equation). Consequently, we still conclude that the value function is C^2 , convex, and that the optimal policy exists in the manner that was discussed in the previous sections.

Note that condition (2.10) now becomes

$$(5.2) \quad \alpha > 2|a| + \theta^2 a^2.$$

We modify condition (2.9) (of assumption (A2)) by

$$(5.3) \quad \left(x + \frac{b}{a}\right) h'(x) \geq 0, \quad x \in \mathbf{R}; \quad h'\left(-\frac{b}{a}\right) = 0;$$

i.e., we restrict the vertex of h to the point $x = -b/a$ so as to “kill” the singularity caused by σ .

Observe that, if we set $Y(t) = Y^{x+b/a}(t) = X^x(t) + b/a$, where $X^x(\cdot)$ is the solution of the Stochastic differential equation (S.D.E.) (2.1) with $\sigma = \theta(ax + b)$, then $Y(\cdot)$ will satisfy

$$(5.4) \quad Y(t) = \tilde{x} + a \int_0^t Y(s) ds + \theta a \int_0^t Y(s) dW(s) + \xi(t),$$

where $\tilde{x} = (x + b/a)$. Therefore, the cost function (2.2) becomes

$$(5.5) \quad V_{\xi}(x) = E \int_{[0, \infty)} e^{\alpha t} [d\tilde{\xi}(t) + h(Y^{\tilde{x}}(t) - \frac{b}{a})dt] \stackrel{\text{def}}{=} \tilde{V}_{\xi}(\tilde{x}).$$

Define $\tilde{h}(\cdot) = h(\cdot - b/a)$; then, by (5.3), \tilde{h} satisfies (2.8) and

$$(5.6) \quad x\tilde{h}'(x) \geq 0, \quad x \in \mathbf{R}; \quad \tilde{h}'(0) = 0.$$

So, without loss of generality, we may just consider system (5.4) with the cost function (5.5). Namely, we will henceforth assume that $b = 0$ and h satisfies (5.6).

The ODE (3.1) now becomes

$$(5.7) \quad \alpha V(x) = axV'(x) + \frac{1}{2}\theta^2 a^2 x^2 V''(x) + h(x).$$

We see that it is now symmetric with respect to the origin and is in the form of the *Euler equation*. So we may solve it explicitly to get the smooth fitting points.

To begin, we first solve the equation for $x > 0$. Letting $U = V'$ and differentiating (5.7), we get that

$$(5.8) \quad \frac{1}{2}\theta^2 a^2 x^2 U''(x) + (a + \theta^2 a^2)xU'(x) + (a - \alpha)U(x) + h'(x) = 0.$$

Set $x = e^t$, $t \in \mathbf{R}$, let $W(t) = U(e^t) = U(x)$, and denote $\dot{W} = dW/dt$, $\ddot{W} = d^2W/dt^2$; (5.8) becomes

$$(5.9) \quad \frac{1}{2}\theta^2 a^2 \ddot{W}(t) + (a + \frac{1}{2}\theta^2 a^2)\dot{W}(t) + (a - \alpha)W(t) + h'(e^t) = 0.$$

We may easily write the solution of (5.9) as

$$(5.10) \quad W(t) = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t} - \frac{2}{\theta^2 a^2} \int_{\ln B}^t \varphi(t - \tau) h'(e^{\tau}) d\tau,$$

where $B > 0$ is arbitrarily chosen and φ is the solution of the homogeneous equation

$$(5.11) \quad \frac{1}{2}\theta^2 a^2 \ddot{W}(t) + (a + \frac{1}{2}\theta^2 a^2)\dot{W}(t) + (a - \alpha)W(t) = 0$$

with the initial condition

$$(5.12) \quad \varphi(0) = 0; \quad \dot{\varphi}(0) = 1,$$

and λ_1, λ_2 are the solutions of the characteristic equation

$$(5.13) \quad \frac{1}{2}\theta^2 a^2 \lambda^2 + (a + \frac{1}{2}\theta^2 a^2)\lambda + (a - \alpha) = 0.$$

Namely,

$$(5.14) \quad \lambda_1 = \frac{-(a + \frac{1}{2}\theta^2 a^2) + \sqrt{(a - \frac{1}{2}\theta^2 a^2)^2 + 2\theta^2 a^2 \alpha}}{\theta^2 a^2};$$

$$(5.15) \quad \lambda_2 = \frac{-(a + \frac{1}{2}\theta^2 a^2) - \sqrt{(a - \frac{1}{2}\theta^2 a^2)^2 + 2\theta^2 a^2 \alpha}}{\theta^2 a^2}.$$

Clearly, if $\alpha > |a|$, then $\lambda_1 > 0 > \lambda_2$, and, if we rewrite (5.13) as

$$(5.16) \quad 0 = [\frac{1}{2}\theta^2 a^2 \lambda^2 + (a + \theta^2 a^2)](\lambda - 1) + (2a + \theta^2 a^2 - \alpha),$$

then it follows from (5.2) and $\lambda_1 > 0$ that $\lambda_1 > 1$. Furthermore, if we let $\varphi(t) = \hat{C}_1 e^{\lambda_1 t} + \hat{C}_2 e^{\lambda_2 t}$, then (5.12) gives

$$(5.17) \quad \hat{C}_1 = -\frac{1}{\lambda_2 - \lambda_1}; \quad \hat{C}_2 = \frac{1}{\lambda_2 - \lambda_1}.$$

Therefore (5.10) becomes

$$(5.18) \quad W(t) = \sum_{i=1}^2 [C_i e^{\lambda_i t} - \frac{2\hat{C}_i}{\theta^2 a^2} \int_{\ln B}^t e^{\lambda_i(t-\tau)} h'(e^\tau) d\tau].$$

In terms of the original variable, i.e., $t = \ln x; x > 0$, we get that, for each $B > 0$,

$$(5.19) \quad U_B(x) = \sum_{i=1}^2 x^{\lambda_i} [C_i - \frac{2\hat{C}_i}{\theta^2 a^2} \int_B^x \frac{h'(u)}{u^{\lambda_i+1}} du].$$

We now choose C_1, C_2 so that $U_B(0+) = 0; U_B(B) = 1$. To do this, we first give the following lemma.

LEMMA 5.1. *For any C^2 -function h satisfying (2.8), (5.6) and λ_1, λ_2 given by (5.14), (5.15), we have that*

- (a) $\lim_{x \rightarrow 0+} x^{\lambda_1} \int_B^x \frac{h'(u)}{u^{\lambda_1+1}} du = 0;$
- (b) $\lim_{x \rightarrow 0+} x^{\lambda_2} \int_0^x \frac{h'(u)}{u^{\lambda_2+1}} du = 0;$
- (c) $\lim_{x \rightarrow 0+} x^{\lambda_1-1} \int_B^x \frac{h'(u)}{u^{\lambda_1+1}} du = \frac{h''(0)}{1-\lambda_1};$
- (d) $\lim_{x \rightarrow 0+} x^{\lambda_2-1} \int_0^x \frac{h'(u)}{u^{\lambda_2+1}} du = \frac{h''(0)}{1-\lambda_2}.$

Proof. (a) and (c). Since $\lambda_1 > 1$, $\lim_{x \rightarrow 0+} x^{-\lambda_1} = \lim_{x \rightarrow 0+} x^{-\lambda_1+1} = +\infty$. Thus, by L'Hospital's rule, we have that

$$\begin{aligned} \lim_{x \rightarrow 0+} x^{\lambda_1} \int_B^x \frac{h'(u)}{u^{\lambda_1+1}} du &= \lim_{x \rightarrow 0+} \frac{h'(x)/x^{\lambda_1+1}}{(-\lambda_1)x^{-\lambda_1-1}} = \frac{h'(0)}{(-\lambda_1)} = 0; \\ \lim_{x \rightarrow 0+} x^{\lambda_1-1} \int_B^x \frac{h'(u)}{u^{\lambda_1+1}} du &= \lim_{x \rightarrow 0+} \frac{h'(x)/x^{\lambda_1+1}}{(1-\lambda_1)x^{-\lambda_1}} = \frac{1}{1-\lambda_1} \lim_{x \rightarrow 0+} \frac{h'(x)}{x}, \\ &= \frac{h''(0)}{1-\lambda_1}, \end{aligned}$$

here we use the fact that $h'(0) = 0$.

(b) and (d). The proof is similar to the previous one, except that now we have that $\lim_{x \rightarrow 0+} x^{-\lambda_2} = \lim_{x \rightarrow 0+} x^{-\lambda_2+1} = 0$, since $\lambda_2 < 0$. So we can apply the previous argument to get the result, provided that we can show that $\lim_{x \rightarrow 0+} \int_0^x (h'(u)/u^{\lambda_2+1}) du = 0$.

Observe that, by integration by parts,

$$(5.20) \quad \int_0^x \frac{h'(u)}{u^{\lambda_2+1}} du = \frac{1}{(-\lambda_2)} \left[\frac{h'(u)}{u^{\lambda_2}} \Big|_0^x - \int_0^x \frac{h''(u)}{u^{\lambda_2}} du \right].$$

So the result follows from $\lambda_2 < 0$. \square

By (5.20) we see that $0 \leq \int_0^B (h'(u)/u^{\lambda_2+1}) du < \infty$. Let $C_2 = -(2\hat{C}_2/\theta^2 a^2) \times \int_0^B (h'(u)/u^{\lambda_2+1}) du$, then (5.19) becomes

$$(5.21) \quad U_B(x) = x^{\lambda_1} \left[C_1 - \frac{2\hat{C}_1}{\theta^2 a^2} \int_B^x \frac{h'(u)}{u^{\lambda_1+1}} du \right] - x^{\lambda_2} \frac{2\hat{C}_2}{\theta^2 a^2} \int_0^x \frac{h'(u)}{u^{\lambda_2+1}} du.$$

Hence Lemma 5.1 (a), (b) imply that $U_B(0+) = 0$. Moreover, by (5.21),

$$U_B(B) = B^{\lambda_1} C_1 - B^{\lambda_2} \frac{2\hat{C}_2}{\theta^2 a^2} \int_0^B \frac{h'(u)}{u^{\lambda_2+1}} du.$$

So $U_B(B) = 1$ if and only if

$$(5.22) \quad C_1 = B^{-\lambda_1} \left[1 + \frac{2B^{\lambda_2} \hat{C}_2}{\theta^2 a^2} \int_0^B \frac{h'(u)}{u^{\lambda_2+1}} du \right].$$

Furthermore, (5.21) also gives

$$\begin{aligned} U'_B(x) &= -\lambda_1 x^{\lambda_1-1} \frac{2\hat{C}_1}{\theta^2 a^2} \int_B^x \frac{h'(u)}{u^{\lambda_1+1}} du + C_1 \lambda_1 x^{\lambda_1-1} \\ &\quad - \lambda_2 x^{\lambda_2-1} \frac{2\hat{C}_2}{\theta^2 a^2} \int_0^x \frac{h'(u)}{u^{\lambda_2+1}} du - \frac{2h'(x)}{\theta^2 a^2 x} [\hat{C}_1 + \hat{C}_2] \\ (5.23) \quad &= -\lambda_1 x^{\lambda_1-1} \frac{2\hat{C}_1}{\theta^2 a^2} \int_B^x \frac{h'(u)}{u^{\lambda_1+1}} du + C_1 \lambda_1 x^{\lambda_1-1} \\ &\quad - \lambda_2 x^{\lambda_2-1} \frac{2\hat{C}_2}{\theta^2 a^2} \int_0^x \frac{h'(u)}{u^{\lambda_2+1}} du, \end{aligned}$$

since $\hat{C}_1 + \hat{C}_2 = 0$.

Using (5.17), (5.23), Lemma 5.1 (c) and (d), and the fact that $\lambda_1 > 1$, we get that

$$\begin{aligned} 0 < U'_B(0+) &= -\lambda_1 \hat{C}_1 \frac{2h''(0)}{\theta^2 a^2 (1 - \lambda_1)} - \lambda_2 \hat{C}_2 \frac{2h''(0)}{\theta^2 a^2 (1 - \lambda_2)} \\ (5.24) \quad &= \frac{2h''(0)[(\lambda_1 \hat{C}_1 + \lambda_2 \hat{C}_2) - \lambda_1 \lambda_2 (\hat{C}_1 + \hat{C}_2)]}{\theta^2 a^2 (\lambda_1 - 1)(1 - \lambda_2)} \quad 5.24 \\ &= \frac{2h''(0)}{\theta^2 a^2 (\lambda_1 - 1)(1 - \lambda_2)} < \infty. \end{aligned}$$

Finally, setting $x = B$ and substituting (5.22) into (5.23), we get that

$$\begin{aligned} U'_B(B) &= \lambda_1 B^{\lambda_1-1} \left\{ B^{-\lambda_1} \left[1 + \frac{2B^{\lambda_2} \hat{C}_2}{\theta^2 a^2} \int_0^B \frac{h'(u)}{u^{\lambda_2+1}} du \right] \right\} \\ (5.25) \quad &= \frac{1}{B} \left\{ \lambda_1 - (\lambda_2 - \lambda_1) B^{\lambda_2} \frac{2\hat{C}_2}{\theta^2 a^2} \int_0^B \frac{h'(u)}{u^{\lambda_2+1}} du \right\} \\ &= \frac{1}{B} \left\{ \lambda_1 - \frac{2B^{\lambda_2}}{\theta^2 a^2} \int_0^B \frac{h'(u)}{u^{\lambda_2+1}} du \right\} \end{aligned}$$

by (5.17). Therefore $U'_B(B) = 0$ if and only if

$$(5.26) \quad \frac{2B^{\lambda_2}}{\theta^2 a^2} \int_0^B \frac{h'(u)}{u^{\lambda_2+1}} du - \lambda_1 = 0.$$

Let $F(B) = (2B^{\lambda_2}/\theta^2 a^2) \int_0^B (h'(u)/u^{\lambda_2+1}) du$. Lemma 5.1(b) gives that $\lim_{B \rightarrow 0+} F(B) = 0$, and the same argument will show that $\lim_{B \rightarrow +\infty} F(B) = +\infty$, since $\lim_{B \rightarrow +\infty} h'(B) = +\infty$ by (2.8). Hence there must be a $B^* > 0$ such that $F(B^*) = \lambda_1$, i.e., $U'_{B^*}(B^*) = 0$ by (5.25).

Since, on $(0, B^*]$, the differential equation (5.7) has no singularity, (5.24) and Lemma 3.2 give that $U'_{B^*}(x) \geq 0$ for all $x \in (0, B^*]$. We now consider that $V_{B^*}(x) = C + \int_0^x U_{B^*}(t) dt$, $0 < x \leq B^*$, where C is some constant. Then V_{B^*} is a solution of (5.7) if and only if $h(0) = \alpha C$. Therefore

$$(5.27) \quad V_{B^*}(x) = \frac{h(0)}{\alpha} + \int_0^x U_{B^*}(t) dt$$

is the solution to (5.7) with the properties

$$(5.28) \quad \begin{aligned} V_{B^*}(0+) &= \frac{h(0)}{\alpha}; \\ V'_{B^*}(0+) &= U_{B^*}(0+) = 0; \quad V'_{B^*}(B^*) = U_{B^*}(B^*) = 1; \\ V''_{B^*}(0+) &= \frac{2h''(0)}{\theta^2 a^2 (\lambda_1 - 1)(1 - \lambda_2)}; \quad V''_{B^*}(B^*) = U'_{B^*}(B^*) = 0; \\ V''_{B^*}(x) &\geq 0, \quad x \in (0, B^*]. \end{aligned}$$

This solves the smooth fitting problem on \mathbf{R}^+ .

To solve (5.7) for $x < 0$, we first consider the following equation:

$$(5.29) \quad \alpha V^1(x) = axV^{1'}(x) + \frac{1}{2}\theta^2 a^2 x^2 V^{1''}(x) + h(-x), \quad x > 0.$$

Conditions (2.8), (5.6) allow us to repeat the previous argument to find a real number $B_1 > 0$ determined by

$$(5.30) \quad \frac{2B_1^{\lambda_2}}{\theta^2 a^2} \int_0^{B_1} \frac{h'(-u)}{u^{\lambda_2+1}} du - \lambda_1 = 0,$$

and a solution $V_{B_1}^1$ to (5.29) for $x > 0$ such that

$$(5.31) \quad \begin{aligned} V_{B_1}^1(0+) &= \frac{h(0)}{\alpha}; \\ V_{B_1}^{1'}(0+) &= 0; \quad V_{B_1}^{1'}(B_1) = 1; \\ V_{B_1}^{1''}(0+) &= \frac{2h''(0)}{\theta^2 a^2 (\lambda_1 - 1)(1 - \lambda_2)}; \quad V_{B_1}^{1''}(B_1) = 0; \\ V_{B_1}^{1''}(x) &\geq 0, \quad x \in (0, B_1]. \end{aligned}$$

We can now define

$$(5.32) \quad V_{L^*, B^*}(x) = \begin{cases} V^1(-x), & x \in [L^*, 0); \\ \frac{h(0)}{\alpha}, & x = 0; \\ V(x), & x \in (0, B^*], \end{cases}$$

where $L^* = -B_1$. It is easily checked, by using (5.27), (5.28), and (5.31), that V_{L^*, B^*} is a C^2 -solution to (5.7) and is convex on $[L^*, B^*]$ satisfying (3.2), (3.3). Therefore Theorem 4.1 applies. We have actually proved the following theorem.

THEOREM 5.2. *For the linear system*

$$X(t) = x + \int_0^t (aX(s) + b)ds + \int_0^t \theta(aX(s) + b)dW(s) + \xi(t),$$

where a, b, θ are constants, $a \neq 0, \theta \neq 0$, there always exist $-\infty < L^* < B^* < \infty$ and an optimal control given by (4.5), provided that (5.2), (5.3) hold. The value function V^* is convex, and C^2 and is given by (4.4) with V_{L^*, B^*} given by (5.32). Moreover, the “smooth fit points” $L^*(= -B_1), B^*$ are determined by (5.30) and (5.26), along with (5.14), (5.15).

6. Appendix. We now outline the proof of Lemma 3.1. We can always refer to [12, Lemmas 4.1, 4.2] for complete details.

Proof of Lemma 3.1. By (2.5), σ is nonvanishing; so we can rewrite (3.4) as

$$(6.1) \quad V''(x) = \gamma(x)V(x) + \delta(x)V'(x),$$

where

$$\gamma(x) = \frac{2\alpha}{\sigma^2(x)}, \quad \delta(x) = -\frac{ax+b}{\sigma^2(x)}.$$

Introduce a change of variable $\phi(x) = \int_0^x [\exp \int_0^u \delta(v)dv]du$ and define $U(y) = V \circ \phi^{-1}(y)$; then U satisfies

$$(6.2) \quad U'''(y) = [\phi'(\phi^{-1}(y))]^{-2} \gamma(\phi^{-1}(y))U(y) = \tilde{\gamma}(y)U(y), \quad y \in [\tilde{L}, \tilde{B}],$$

where $\tilde{\gamma}(y) = [\phi'(\phi^{-1}(y))]^{-2} \gamma(\phi^{-1}(y)) > 0$; $\tilde{L} = \phi(L)$, $\tilde{B} = \phi(B)$. It is readily seen that U (respectively, U') and V (respectively, V') have the same sign; hence V inherits the desired properties from U . Namely, without loss of generality, we may assume that $a = b = 0$, and then (6.1) becomes

$$(6.3) \quad V''(x) = \gamma(x)V(x), \quad x \in [L, B],$$

where $\gamma > 0$. Observe now that V is strictly convex (strictly concave) on any interval where it is positive (negative).

To prove (i), suppose that $V(\bar{x}) = 0$, for some $\bar{x} \in [L, B]$. Then we must have that $V'(\bar{x}) \neq 0$; otherwise, $V \equiv 0$ by the uniqueness of the solution to (6.3). Suppose that $V'(\bar{x}) > 0$. Define

$$\bar{w} = \sup \{w \in [\bar{x}, B] : V'(x) > 0, \text{ for } \bar{x} \leq x < w\}.$$

By a simple analysis on the signs of V' , V , and V'' on the interval (\bar{x}, \bar{w}) , we show that V is convex on (\bar{x}, \bar{w}) , which implies that $V'(\bar{w}) \geq V'(\bar{x}) > 0$. Then the definition of \bar{w} and the continuity of V' lead to $\bar{w} = B$. Therefore V' has no zero on $[\bar{x}, B]$. Similarly, we can show that V' has no zero on $[L, \bar{x}]$. The case where $V'(\bar{x}) < 0$ is treated similarly. This proves (i).

To prove (ii), assume that $V'(\bar{x}) = 0$ for some $\bar{x} \in [L, B]$. Again, we must have that $V(\bar{x}) \neq 0$. Without loss of generality, assume that $V(\bar{x}) > 0$. By (i), V would

have no zero on $[L, B]$ as V' has a zero at $\bar{x} \in [L, B]$. So $V(x) > 0$ for all $x \in [L, B]$; i.e., V is strictly convex on $[L, B]$, which implies that

$$(x - \bar{x})V'(x) > 0, \quad \text{for all } x \in [L, B], \quad x \neq \bar{x},$$

since $V'(\bar{x}) = 0$. Then

$$(x - \bar{x})V'(x)V(x) > 0, \quad x \in [L, B]. \quad \square$$

Acknowledgments. I would like to express my deep gratitude to my advisor Professor Naresh Jain for many helpful discussions. I also thank Professors S. E. Shreve and I. Karatzas for providing information on recent research in this area, along with their papers. Finally, I thank the referees for their useful comments.

REFERENCES

- [1] V. E. BENÈS, L. A. SHEPP, AND H. S. WITSENHAUSEN, *Some solvable stochastic control problems*, *Stochastics*, 4 (1980), pp. 39–83.
- [2] P. L. CHOW, J. L. MENALDI, AND M. ROBIN, *Additive control of stochastic linear systems with finite horizon*, *SIAM. J. Control Optim.*, 23 (1985), pp. 858–899.
- [3] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw–Hill, New York, Toronto, London, 1955.
- [4] A. FRIEDMANN, *Partial Differential Equations of Parabolic Type*, Prentice–Hall, Englewood Cliffs, NJ, 1965.
- [5] I. I. GIKHMAN AND A. V. SKOROHOD, *Stochastic Differential Equations*, Springer, New York, 1972.
- [6] J. M. HARRISON AND M. I. TAKSAR, *Instantaneous control of Brownian motion*, *Math. Oper. Res.*, 8 (1983), pp. 439–453.
- [7] U. G. HAUSSMANN, *A Stochastic Maximum Principle for Optimal control of Diffusions*, Pitman Research Notes in Math. Series, Longman Scientific & Technical, London, 1986.
- [8] I. KARATZAS, *A class of singular stochastic control problems*, *Adv. Appl. Probab.*, 15 (1983), pp. 225–254.
- [9] J. L. MENALDI AND M. ROBIN, *On some cheap control problems for diffusion processes*, *Trans. Amer. Math. Soc.*, 278 (1983), pp. 771–802.
- [10] J. L. MENALDI AND M. I. TAKSAR, *Optimal correction problem of a multidimensional stochastic system*, *Automatica*, 25 (1989), pp. 223–232.
- [11] P. A. MEYER, *Un cours sur les intégral stochastiques*, *Séminaire de Probabilité, X.*, Lecture Notes in Math., 511, Springer, New York, 1976.
- [12] S. E. SHREVE, *An Introduction To Singular Stochastic Control*, in *Stochastic Differential Systems, Stochastic Control Theory and Applications*, IMA Volumes in Math and Its Applications, 10, Springer, New York, 1986.
- [13] S. E. SHREVE, J. P. LEHOCZKY, AND D. P. GAVER, *Optimal consumption for general diffusions with absorbing and reflecting barriers*, *SIAM. J. Control Optim.*, 22 (1984), pp. 55–75.
- [14] H. M. SONER AND S. E. SHREVE, *Regularity of the value function for a two-dimensional singular stochastic control problem*, *SIAM. J. Contr. Optim.*, 27 (1989), pp. 876–907.

EXACT SEMI-INTERNAL CONTROL OF AN EULER-BERNOULLI EQUATION*

JONG UHN KIM†

Abstract. This paper proves exact semi-internal controllability of an Euler-Bernoulli equation with variable coefficients. The basic principle of the proof is the Hilbert uniqueness method. For this, a multiplier technique and a unique continuation property have been used. The proof of the unique continuation property is the main feature of this work.

Key words. exact controllability, Euler-Bernoulli equation, Hilbert uniqueness method, Carleman estimate, unique continuation property

AMS(MOS) subject classifications. 35Q20, 35B37, 35B60, 49E15

Introduction. The purpose of this paper is to prove exact semi-internal controllability of an Euler-Bernoulli equation with the following variable coefficients:

$$(0.1) \quad u_{tt} + a(t)\Delta^2 u + \sum_{|\alpha| \geq 2} b_\alpha(x, t) \partial_x^\alpha u + d(x, t)u_t = f(x, t) \quad \text{in } \Omega \times (0, T),$$

$$(0.2) \quad u = \frac{\partial u}{\partial \nu} = 0 \quad \text{on } \partial\Omega \times (0, T),$$

$$(0.3) \quad u(x, 0) = u_0(x), \quad u_t(x, 0) = u_1(x) \quad \text{in } \Omega.$$

Here Ω is a bounded open subset of R^n with smooth boundary, and $\partial/\partial\nu$ stands for the normal derivative on $\partial\Omega$. We use the notation $\partial_t = \partial/\partial t$, $\partial_x^\alpha = (\partial/\partial x_1)^{\alpha_1} \cdots (\partial/\partial x_n)^{\alpha_n}$, $\alpha = (\alpha_1, \dots, \alpha_n)$, and $|\alpha| = \alpha_1 + \cdots + \alpha_n$. $f(x, t)$ denotes a control that is required to be supported in $E \times (0, T)$, where E is a given neighborhood of the whole boundary $\partial\Omega$.

The problem of exact controllability is to find a control $f(x, t)$ that can drive the solution of (0.1)–(0.3) to a desired final state $(\tilde{u}_0, \tilde{u}_1)$ at a given time. Equation (0.1) is a variant of

$$(0.4) \quad u_{tt} + \Delta^2 u = f,$$

which describes the motion of a homogeneous elastic plate. For various plate models, see Lagnese [9]. Zuazua [14] proved exact semi-internal controllability of (0.4) with a control f supported in $\tilde{E} \times (0, T)$, where \tilde{E} is a neighborhood of a certain subset of $\partial\Omega$. If the support of control is required to be strictly in the interior of Ω , the problem is still open even for the simple equation (0.4) when $n \geq 2$. Only the special case where Ω is a rectangle has been resolved; see Haraux [2], Jaffard [4], [5], and Komornik [8]. The general one-dimensional equation with time-independent variable coefficients was discussed by the author [7] without restriction on the location of the control. When $a(t) \equiv \text{constant}$, and the b_α 's and d 's are time independent, the known argument can be still used with aid of a unique continuation property based on the argument of Bardos, Lebeau, and Rauch [1].

The basic principle in the proof of exact controllability is the Hilbert uniqueness method (HUM) due to Lions [12]. For this, we typically employ a multiplier technique and a unique continuation property. When the coefficients are variable, a standard

* Received by the editors July 2, 1990; accepted for publication (in revised form) May 5, 1991.

† Department of Mathematics, Virginia Polytechnic and State University, Blacksburg, Virginia 24061. This research was supported by Air Force Office of Scientific Research grant AFOSR-89-0268.

multiplier can be still used, but the unique continuation property requires a new proof. First, Holmgren's uniqueness theorem is out of consideration since we do not assume the analyticity of coefficients. Second, the argument of Bardos, Lebeau, and Rauch [1] cannot be used with coefficients that depend on both the time and space variables. Hence, the main task of this paper is to establish a unique continuation property. Our proof is based upon the Carleman estimate for a Schrödinger equation due to Isakov [3], which was also used by Lasiecka and Triggiani [11]. Similar estimates for quasi-homogeneous equations were also obtained by Khalgui-Ounaies [6] and Zuily [15]. Since the Carleman estimates obtained in these works are of a local nature and require a strong pseudoconvexity condition, technical complexity arises in proving a global unique continuation property by means of such estimates. We resolve this difficulty by an elementary geometric reduction.

In § 1, we state the result on exact controllability and outline how a unique continuation property yields exact controllability. In § 2, we present a proof of the unique continuation property, which is the main feature of this work.

Notation. When f is a locally integrable function in a domain $G \subset R^{n+1}$, $\text{supp } f$ denotes the smallest closed subset of G in the complement of which f is almost everywhere equal to zero. We say that $f = 0$ in G if f is zero almost everywhere in G .

1. Statement of exact controllability. Let Ω be a bounded open subset of R^n with smooth boundary $\partial\Omega$. Let $E = \Omega \cap \mathcal{O}$, where \mathcal{O} is an open neighborhood of $\partial\Omega$ in R^n . We fix any $T > 0$ and assume that

$$(1.1) \quad a(t) \in C^2([0, T]), \quad a(t) > 0 \quad \text{for all } t \in [0, T],$$

$$(1.2) \quad \begin{aligned} b_\alpha(x, t) &\in C(\bar{\Omega} \times [0, T]) \quad \text{for } |\alpha| \leq 2; \\ \partial_x^\beta b_\alpha(x, t) &\in L^\infty(\Omega \times (0, T)) \quad \text{for } |\alpha| \leq 2, \quad |\beta| \leq |\alpha| + 1, \\ \partial_t b_\alpha(x, t) &\in L^\infty(\Omega \times (0, T)) \quad \text{for } |\alpha| \leq 2; \end{aligned}$$

$$(1.3) \quad \begin{aligned} d(x, t) &\in C(\bar{\Omega} \times [0, T]), \\ \partial_x^\beta d(x, t) &\in L^\infty(\Omega \times (0, T)) \quad \text{for } |\beta| \leq 3, \\ \partial_t \partial_x^\gamma d(x, t) &\in L^\infty(\Omega \times (0, T)) \quad \text{for } |\gamma| \leq 1. \end{aligned}$$

THEOREM 1.1. *For given $u_0 \in H_0^2(\Omega)$ and $u_1 \in L^2(\Omega)$, there is $f(x, t) \in L^2(\Omega \times (0, T))$ with $\text{supp } f \subset \bar{E} \times (0, T)$ such that the solution of (0.1)–(0.3) satisfies*

$$(1.4) \quad u(x, T) = 0, \quad u_t(x, T) = 0 \quad \text{in } \Omega.$$

Since we do not impose any sign condition on $d(x, t)$, this null-controllability implies exact controllability. By a transformation in the time variable, we can put (0.1) in a simpler form. Let us define

$$(1.5) \quad p(t) = \int_0^t (a(\sigma))^{1/2} d\sigma$$

and set

$$(1.6) \quad s = p(t),$$

$$(1.7) \quad \tilde{u}(x, s) = u(x, q(s)),$$

where $q(\cdot)$ is the inverse of $p(\cdot)$, i.e., $t = q(s)$. Then, (0.1) is equivalent to

$$(1.8) \quad \tilde{u}_{ss} + \Delta^2 \tilde{u} + \sum_{|\alpha| \leq 2} \tilde{b}_\alpha(x, s) \partial_x^\alpha \tilde{u} + \tilde{d}(x, s) \tilde{u}_s = \tilde{f}(x, s),$$

where

$$\begin{aligned}\tilde{b}_\alpha(x, s) &= b_\alpha(x, q(s))/a(q(s)) \quad \text{for } |\alpha| \leq 2, \\ \tilde{d}(x, s) &= (d(x, q(s))p'(q(s)) + p''(q(s)))/a(q(s)), \\ \tilde{f}(x, s) &= f(x, q(s))/a(q(s)).\end{aligned}$$

By virtue of (1.1)–(1.3), we have

$$\begin{aligned}(1.9) \quad & \tilde{b}_\alpha(x, s) \in C(\bar{\Omega} \times [0, \tilde{T}]) \quad \text{for } |\alpha| \leq 2, \\ & \partial_x^\beta \tilde{b}_\alpha(x, s) \in L^\infty(\Omega \times (0, \tilde{T})) \quad \text{for } |\alpha| \leq 2, \quad |\beta| \leq |\alpha| + 1, \\ & \partial_s \tilde{b}_\alpha(x, s) \in L^\infty(\Omega \times (0, \tilde{T})) \quad \text{for } |\alpha| \leq 2; \\ (1.10) \quad & \tilde{d}(x, s) \in C(\bar{\Omega} \times [0, \tilde{T}]), \\ & \partial_x^\beta \tilde{d}(x, s) \in L^\infty(\Omega \times (0, \tilde{T})), \quad |\beta| \leq 3, \\ & \partial_s \partial_x^\gamma \tilde{d}(x, s) \in L^\infty(\Omega \times (0, \tilde{T})), \quad |\gamma| \leq 1,\end{aligned}$$

where $\tilde{T} = p(T)$.

It is obvious that Theorem 1.1 follows if we prove null-controllability of (1.8). We rewrite (1.8), by replacing s by t and suppressing the tilde, as

$$(1.11) \quad u_{tt} + \Delta^2 u + \sum_{|\alpha| \leq 2} b_\alpha(x, t) \partial_x^\alpha u + d(x, t) u_t = f(x, t) \quad \text{in } \Omega \times (0, T),$$

where b_α and d satisfy (1.2) and (1.3).

THEOREM 1.2. *For given $u_0 \in H_0^2(\Omega)$ and $u_1 \in L^2(\Omega)$, there is $f(x, t) \in L^2(\Omega \times (0, T))$ with $\text{supp } f \subset \bar{E} \times (0, T)$ such that the solution of (1.11), (0.2), and (0.3) satisfies*

$$(1.12) \quad u(x, T) = 0, \quad u_t(x, T) = 0 \quad \text{in } \Omega.$$

Proof. We must consider the following dual problem:

$$(1.13) \quad v_{tt} + \Delta^2 v + \sum_{|\alpha| \leq 2} (-1)^{|\alpha|} \partial_x^\alpha (b_\alpha(x, t) v) - \partial_t (d(x, t) v) = 0 \quad \text{in } \Omega \times (0, T),$$

$$(1.14) \quad v = \frac{\partial v}{\partial \nu} = 0 \quad \text{on } \partial\Omega \times (0, T),$$

$$(1.15) \quad v(x, 0) = v_0(x), \quad v_t(x, 0) = v_1(x) \quad \text{in } \Omega.$$

Let $(v_0, v_1) \in H_0^2(\Omega) \times L^2(\Omega)$. Then, there is a unique solution v in $C([0, T]; H_0^2(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ of (1.13)–(1.15). Next, we let $f = \chi_E v$ in (1.11) with χ_E = the characteristic function of E , and u be a unique solution in $C([0, T]; H_0^2(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ of (1.11), (0.2), and

$$(1.16) \quad u(x, T) = 0, \quad u_t(x, T) = 0 \quad \text{in } \Omega.$$

We then define a mapping Λ by

$$(1.17) \quad \Lambda((v_0, v_1)) = (u(x, 0), u_t(x, 0)).$$

LEMMA 1.3. *For every $(v_0, v_1) \in H_0^2(\Omega) \times L^2(\Omega)$, we have*

$$(1.18) \quad \|\Lambda((v_0, v_1))\|_{H_0^2(\Omega) \times L^2(\Omega)} \leq M \|(v_0, v_1)\|_{L^2(\Omega) \times H^{-2}(\Omega)},$$

with some positive constant M .

Proof. The proof of Lemma 1.3 is postponed.

Hence, Λ can be extended as a continuous linear mapping from $L^2(\Omega) \times H^{-2}(\Omega)$ into $H_0^2(\Omega) \times L^2(\Omega)$. Next, we define a continuous bilinear functional $B(\cdot, \cdot)$ on $(H_0^2(\Omega) \times L^2(\Omega)) \times (L^2(\Omega) \times H^{-2}(\Omega))$ by

$$(1.19) \quad B((v_0, v_1), (w_0, w_1)) = - \int_{\Omega} (v_1 w_0 + d(x, 0) v_0 w_0) dx + \langle v_0, w_1 \rangle,$$

for each $(v_0, v_1) \in H_0^2(\Omega) \times L^2(\Omega)$ and $(w_0, w_1) \in L^2(\Omega) \times H^{-2}(\Omega)$, where $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $H_0^2(\Omega)$ and $H^{-2}(\Omega)$. Obviously, $B(\Lambda(\cdot), \cdot)$ is a continuous bilinear functional on $(L^2(\Omega) \times H^{-2}(\Omega))^2$. When $(v_0, v_1) \in H_0^2(\Omega) \times L^2(\Omega)$ and $f = \chi_E v$ in (1.11), we multiply (1.11) and (1.13) by v and u , respectively, and integrate by parts over $\Omega \times (0, T)$, using (1.16) to obtain

$$(1.20) \quad B(\Lambda(v_0, v_1), (v_0, v_1)) = \int_0^T \int_E v^2 dx dt.$$

This procedure can be justified since (u, u_t) and (v, v_t) belong to $C([0, T]; H_0^2(\Omega) \times L^2(\Omega))$.

PROPOSITION 1.4. *Let v be a unique solution in $C([0, T]; H_0^2(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ of (1.13)–(1.15). Then it holds that*

$$(1.21) \quad \int_0^T \int_E v^2 dx dt \cong M(\|v_0\|_{L^2(\Omega)}^2 + \|v_1\|_{H^{-2}(\Omega)}^2)$$

for some positive constant M independent of (v_0, v_1) .

Proof. The proof of Proposition 1.4 is postponed.

By virtue of (1.20) and (1.21), we can use the Lax–Milgram lemma to find a unique $(v_0, v_1) \in L^2(\Omega) \times H^{-2}(\Omega)$ for given $(u_0, u_1) \in H_0^2(\Omega) \times L^2(\Omega)$ such that

$$(1.22) \quad B((u_0, u_1), (w_0, w_1)) = B(\Lambda(v_0, v_1), (w_0, w_1))$$

for every $(w_0, w_1) \in L^2(\Omega) \times H^{-2}(\Omega)$.

Consequently, we have

$$(1.23) \quad \Lambda(v_0, v_1) = (u_0, u_1).$$

For given (u_0, u_1) in Theorem 1.2, there is $(v_0, v_1) \in L^2(\Omega) \times H^{-2}(\Omega)$ satisfying (1.23), and a sequence $\{(v_0^m, v_1^m)\}_{m=1}^\infty$ in $H_0^2(\Omega) \times L^2(\Omega)$ that converges to (v_0, v_1) in $L^2(\Omega) \times H^{-2}(\Omega)$. Then, the desired control $f(x, t)$ is given by

$$(1.24) \quad f = \lim_{m \rightarrow \infty} \chi_E v^m \quad \text{in } L^2(\Omega \times (0, T)),$$

where v^m is a unique solution in $C([0, T]; H_0^2(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ of (1.13), (1.14), and

$$(1.25) \quad v^m(x, 0) = v_0^m, \quad v_t^m(x, 0) = v_1^m.$$

Here, the limit in (1.24) exists by (1.32) below.

It now remains to prove Lemma 1.3 and Proposition 1.4.

1.1. Proof of Lemma 1.3. The idea is borrowed from Lions [12]. We set

$$(1.26) \quad w(x, t) = \int_0^t v(x, t) dt + \chi(x),$$

where v is a unique solution in $C([0, T]; H_0^2(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ of (1.13)–(1.15), and $\chi(x)$ is a unique solution in $H_0^2(\Omega)$ of

$$(1.27) \quad \Delta^2 \chi + \sum_{|\alpha| \leq 2} (-1)^{|\alpha|} \partial_x^\alpha (b_\alpha(x, 0) \chi) + K \chi = -v_1 + d(x, 0) v_0,$$

where K is a sufficiently large positive number so that (1.27) is uniquely solvable. Then, it is easy to see that w defined by (1.26) is a unique solution in $C([0, T]; H_0^2(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ of

$$(1.28) \quad \begin{aligned} w_{tt} + \Delta^2 w + \sum_{|\alpha| \leq 2} (-1)^{|\alpha|} \partial_x^\alpha (b_\alpha(x, t) w) - d(x, t) w_t + K w \\ - \int_0^t \left\{ K w_t(x, \sigma) + \sum_{|\alpha| \leq 2} (-1)^{|\alpha|} \partial_x^\alpha (b_{\alpha t}(x, \sigma) w(x, \sigma)) \right\} d\sigma = 0, \end{aligned}$$

$$(1.29) \quad w = \frac{\partial w}{\partial \nu} = 0 \quad \text{on } \partial\Omega \times (0, T),$$

$$(1.30) \quad w(x, 0) = \chi(x), \quad w_t(x, 0) = v_0(x) \quad \text{in } \Omega.$$

For the regularity and uniqueness of solutions, see [13]. Therefore, it follows that

$$(1.31) \quad \int_0^T \int_E w_t^2 dx dt \leq M(\|\chi\|_{H_0^2(\Omega)}^2 + \|v_0\|_{L^2(\Omega)}^2),$$

which implies

$$(1.32) \quad \int_0^T \int_E v^2 dx dt \leq M(\|v_0\|_{L^2(\Omega)}^2 + \|v_1\|_{H^{-2}(\Omega)}^2)$$

for some positive constant M independent of (v_0, v_1) . Combining this and

$$(1.33) \quad \begin{aligned} \|u(x, 0)\|_{H_0^2(\Omega)}^2 + \|u_t(x, 0)\|_{L^2(\Omega)}^2 &\leq M \int_0^T \int_\Omega f^2 dx dt \\ &= M \int_0^T \int_E v^2 dx dt, \end{aligned}$$

we obtain (1.18).

1.2. Proof of Proposition 1.4. According to the above context, it is enough to show that

$$(1.34) \quad \int_0^T \int_E w_t^2 dx dt \leq M(\|w_0\|_{H_0^2(\Omega)}^2 + \|w_1\|_{L^2(\Omega)}^2),$$

where w is a unique solution in $C([0, T]; H_0^2(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ of (1.28), (1.29), and

$$(1.35) \quad w(x, 0) = w_0(x), \quad w_t(x, 0) = w_1(x) \quad \text{in } \Omega.$$

Assume that (1.34) is false. Then, there is a sequence $\{(w_0^m, w_1^m)\}_{m=1}^\infty$ such that

$$(1.36) \quad \|w_0^m\|_{H_0^2(\Omega)}^2 + \|w_1^m\|_{L^2(\Omega)}^2 = 1,$$

for each m and the corresponding $w^m(x, t)$ satisfies

$$(1.37) \quad \lim_{m \rightarrow \infty} \int_0^T \int_E (w_t^m)^2 dx dt = 0.$$

By means of (1.36) and (1.37), we can extract a subsequence still denoted by $\{w^m\}_{m=1}^\infty$ such that

$$(1.38) \quad w^m \rightarrow w^\infty \quad \text{weak * in } L^\infty(0, T; H_0^2(\Omega)),$$

$$(1.39) \quad w_t^m \rightarrow w_t^\infty \quad \text{weak * in } L^\infty(0, T; L^2(\Omega)),$$

for some $w^\infty \in C([0, T]; H_0^2(\Omega)) \cap C^1([0, T]; L^2(\Omega))$, which satisfies (1.28). Furthermore,

$$(1.40) \quad \int_0^T \int_E (w_t^\infty)^2 dx dt = 0.$$

In fact, $w^\infty = 0$ in $\Omega \times (0, T)$ according to the following fact.

PROPOSITION 1.5. *Let $w \in C([0, T]; H_0^2(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ be a solution of (1.28). If $w_t = 0$ in $E \times (0, T)$, then $w = 0$ in $\Omega \times (0, T)$.*

Proof. The proof of Proposition 1.5 is postponed.

Thus, we have

$$(1.41) \quad w^m \rightarrow 0 \quad \text{weak * in } L^\infty(0, T; H_0^2(\Omega)),$$

$$(1.42) \quad w_t^m \rightarrow 0 \quad \text{weak * in } L^\infty(0, T; L^2(\Omega)),$$

from which it follows that

$$(1.43) \quad w^m \rightarrow 0 \quad \text{strongly in } C([0, T]; H_0^1(\Omega)).$$

It also follows that each w^m satisfies

$$(1.44) \quad w_{tt}^m + \Delta^2 w^m + G^m(x, t) = 0,$$

where

$$(1.45) \quad G^m \rightarrow 0 \quad \text{weak * in } L^\infty(0, T; L^2(\Omega)).$$

Next, we let \mathcal{O}_0 be an open subset of R^n such that $\partial\Omega \subset \mathcal{O}_0 \subset \bar{\mathcal{O}}_0 \subset \mathcal{O}$. Recall that $E = \Omega \cap \mathcal{O}$ and set $E_0 = \Omega \cap \mathcal{O}_0$. Choose a nonnegative function $\phi \in C_0^\infty(R^n)$ such that $\phi = 1$ on $\bar{\mathcal{O}}_0$ and $\text{supp } \phi \cap \Omega \subset E$. Multiply (1.44) by ϕw^m and integrate over $\Omega \times (0, T)$ to obtain by (1.37), (1.41)–(1.43), and (1.45)

$$(1.46) \quad \lim_{m \rightarrow \infty} \int_0^T \int_{E_0} (\partial_j \partial_k w^m)^2 dx dt = 0 \quad \text{for } j, k = 1, \dots, n.$$

Next, choose $\psi \in C_0^\infty(\Omega)$ such that $\psi = 1$ on $\Omega \setminus E_0$ and set

$$(1.47) \quad q_k(x) = x_k \psi(x), \quad k = 1, \dots, n, \quad x = (x_1, \dots, x_n).$$

Then, by slightly modifying identity (3.15) on p. 244 of [12], we get

$$(1.48) \quad \begin{aligned} 0 &= \sum_{k=1}^n \int_\Omega w_t^m q_k \partial_k w^m dx \Big|_{t=0}^{t=T} + \frac{1}{2} \sum_{k=1}^n \int_0^T \int_\Omega \{(w_t^m)^2 - (\Delta w^m)^2\} \partial_k q_k dx dt \\ &+ 2 \sum_{j,k=1}^n \int_0^T \int_\Omega \partial_j q_k (\Delta w^m) \partial_j \partial_k w^m dx dt \\ &+ \sum_{k=1}^n \int_0^T \int_\Omega \Delta q_k \Delta w^m \partial_k w^m dx dt + \sum_{k=1}^n \int_0^T \int_\Omega G^m q_k \partial_k w^m dx dt, \end{aligned}$$

which can be rewritten as

$$(1.49) \quad \begin{aligned} 0 &= \frac{n}{2} \int_0^T \int_{\Omega \setminus E_0} \{(w_t^m)^2 - (\Delta w^m)^2\} dx dt \\ &+ 2 \int_0^T \int_{\Omega \setminus E_0} (\Delta w^m)^2 dx dt + H^m \quad \text{for each } m, \end{aligned}$$

where H^m is an obvious remainder.

By virtue of (1.37), (1.41)–(1.43), (1.45), and (1.46), it is evident that

$$(1.50) \quad \lim_{m \rightarrow \infty} H^m = 0.$$

Next, we multiply (1.44) by ψw^m and integrate over $\Omega \times (0, T)$ to obtain

$$(1.51) \quad 0 = \int_0^T \int_{\Omega \setminus E_0} \{-(w_t^m)^2 + (\Delta w^m)^2\} dx dt + F^m \quad \text{for each } m,$$

where F^m is an obvious remainder such that

$$(1.52) \quad \lim_{m \rightarrow \infty} F^m = 0$$

on account of (1.37), (1.41)–(1.43), and (1.45). Combining (1.46), (1.49), and (1.51), we find that

$$(1.53) \quad \lim_{m \rightarrow \infty} \int_0^T \int_{\Omega} (\Delta w^m)^2 dx dt = 0,$$

which, together with (1.37), yields

$$(1.54) \quad \lim_{m \rightarrow \infty} \int_0^T \int_{\Omega} (w_t^m)^2 dx dt = 0.$$

This contradicts (1.36). Now the proof of Proposition 1.4 will be complete if we prove Proposition 1.5, which follows from the next proposition.

PROPOSITION 1.6. *Let $v \in C([0, T]; L^2(\Omega)) \cap C^1([0, T]; H^{-2}(\Omega))$ satisfy (1.13). If $v = 0$ in $E \times (0, T)$, then $v = 0$ in $\Omega \times (0, T)$.*

We show that Proposition 1.5 follows from Proposition 1.6. Let $v = w_t$. Then, $v \in C([0, T]; L^2(\Omega)) \cap C^1([0, T]; H^{-2}(\Omega))$ satisfies (1.13), which is obtained by differentiating (1.28) in t . Furthermore, $v = 0$ in $E \times (0, T)$. According to Proposition 1.6, $v = 0$ in $\Omega \times (0, T)$. Hence, w is independent of t . It follows from (1.28) that $w = \chi(x) \in H_0^2(\Omega)$ satisfies

$$(1.55) \quad \Delta^2 \chi + \sum_{|\alpha| \leq 2} (-1)^{|\alpha|} \partial_x^\alpha (b_\alpha(x, 0) \chi) + K \chi = 0 \quad \text{in } \Omega.$$

Thus, $\chi = 0$ in Ω , which proves Proposition 1.5.

The proof of Proposition 1.6 is presented in § 3, which completes the proof of Theorem 1.2.

2. A unique continuation property. We consider

$$(2.1) \quad v_{tt} + \Delta^2 v + \sum_{|\alpha| \leq 2} B_\alpha(x, t) \partial_x^\alpha v + D(x, t) v_t = 0 \quad \text{in } \Omega \times (0, T),$$

where we assume that

$$(2.2) \quad B_\alpha(x, t) \in L^\infty(\Omega \times (0, T)) \quad \text{for } |\alpha| = 2;$$

$$(2.3) \quad B_\alpha(x, t) \in L^p(\Omega \times (0, T)), \quad p > 2n \quad \text{for } |\alpha| \leq 1;$$

$$(2.4) \quad D(x, t) \in L^\infty(\Omega \times (0, T)).$$

THEOREM 2.1. *Let v be a solution of (2.1) in $\Omega \times (0, T)$ and assume that $v \in L^2(0, T; H^3(\Omega))$ and $v_t \in L^2(0, T; H^1(\Omega))$. If $v = 0$ in $E \times (0, T)$, then $v = 0$ in $\Omega \times (0, T)$.*

Recall that $E = \mathcal{O} \cap \Omega$, where \mathcal{O} is an open subset of R^n that contains $\partial\Omega$.

2.1. Geometric reduction. Let (y, s) be the standard Cartesian coordinate for R^{n+1} with $y \in R^n$, $s \in R$. Denote by $B_r((y^*, s^*))$ an n -dimensional open ball that lies on the hyperplane $s = s^*$ with radius r and center at (y^*, s^*) . Let $(y^{**}, s^*) \in \partial B_r((y^*, s^*))$ and define

$$(2.5) \quad \begin{aligned} & \mathcal{S}((y^*, s^*), r; (y^{**}, s^*), \varepsilon) \\ &= \{(y, s^*): (y, s^*) \in \partial B_r((y^*, s^*)) \text{ and } |y - y^{**}| \leq \varepsilon\}. \end{aligned}$$

Then, $\mathcal{S}((y^*, s^*), r; (y^{**}, s^*), \varepsilon)$ is a closed subset of $\partial B_r((y^*, s^*))$ with center at (y^{**}, s^*) . It is a closed circular arc when $n = 2$. We need the following fact.

PROPOSITION 2.2. *Let $v(x, t)$ be a function defined in $\Omega \times (0, T)$. Suppose that $\text{supp } v$ is not empty and that $\text{supp } v \cap (E \times (0, T))$ is empty. Then, there are $t_0 \in (0, T)$, $x_0 \in \Omega \setminus E$, $y^* \in R^n$ with $r = |y^*| > 0$, positive numbers ε_i , $i = 1, 2, 3$, and a coordinate transformation from (x, t) -space into (y, s) -space as follows:*

$$(2.6) \quad s = t - t_0, \quad y = x - x_0 + h(t),$$

where h is Lipschitz and $h(t_0) = 0$, such that

$$(2.7) \quad w(y, s) = v(x, t)$$

is defined in a neighborhood Q of $y = 0$, $s = 0$ in (y, s) -space,

$$(2.8) \quad \bigcup_{\substack{|\xi| \leq \varepsilon_3 \\ |s| \leq 2\varepsilon_2}} \mathcal{S}(((1 + \xi)y^*, s), r; (\xi y^*, s), \varepsilon_1) \subset Q,$$

$$(2.9) \quad w = 0 \text{ in } \bigcup_{\substack{|\xi| \leq \varepsilon_3 \\ \varepsilon_2 \leq |s| \leq 2\varepsilon_2}} \mathcal{S}(((1 + \xi)y^*, s), r; (\xi y^*, s), \varepsilon_1),$$

$$(2.10) \quad w = 0 \text{ in } \bigcup_{\substack{-\varepsilon_3 \leq \xi < 0 \\ |s| \leq 2\varepsilon_2}} \mathcal{S}(((1 + \xi)y^*, s), r; (\xi y^*, s), \varepsilon_1),$$

$$(2.11) \quad (0, 0) \in \text{supp } w.$$

Proof. Since $\text{supp } v$ is not empty, there is $0 < t^* < T$ such that $\mathcal{G} \stackrel{\text{def}}{=} \text{supp } v \cap (\Omega \times \{t^*\})$ is not empty. Then, $\mathcal{G} \subset (\Omega \setminus E) \times \{t^*\}$ and there are $p \in R^n$ and $r > 0$ such that

$$(2.12) \quad \mathcal{G} \subset \overline{B_r((p, t^*))},$$

$$(2.13) \quad \mathcal{G} \cap \partial B_r((p, t^*)) = \{(x^*, t^*)\} \text{ for some } x^* \in \Omega \setminus E.$$

As in (2.5), we define $\mathcal{S}((p, t^*), r; (x^*, t^*), \varepsilon)$ in (x, t) -space for each $0 < \varepsilon < r$. Next, we can choose $q \in E$ such that (p, t^*) , (x^*, t^*) , and (q, t^*) lie on the same line, and the line segment between (x^*, t^*) and (q, t^*) is contained in $\Omega \times (0, T)$. We then construct a family of uniform \mathcal{S} 's by translating the center of \mathcal{S} along the straight line connecting (x^*, t^*) and (q, t^*) . A typical member of this family is described by

$$\mathcal{S}((p + \xi(q - x^*), t^*), r; (x^* + \xi(q - x^*), t^*), \varepsilon)$$

for some $0 \leq \xi \leq 1$. Refer to Fig. 1 when $n = 2$. Using (2.12) and (2.13), we can choose $0 < \varepsilon < r$, $\xi_0 > 0$, and $\delta_1 > 0$, which are so small that, for all $-\xi_0 \leq \xi \leq 1$ and $|t - t^*| \leq \delta_1$,

$$(2.14) \quad \mathcal{S}((p + \xi(q - x^*), t), r; (x^* + \xi(q - x^*), t), \varepsilon) \subset \Omega \times (0, T)$$

and, for each $0 < \xi \leq 1$,

$$(2.15) \quad \mathcal{S}((p + \xi(q - x^*), t^*), r; (x^* + \xi(q - x^*), t^*), \varepsilon) \subset \Omega \times (0, T) \setminus \text{supp } v.$$

Consequently, there is $0 < \delta_2 \leq \delta_1$ such that for all $|t - t^*| \leq \delta_2$ and $\frac{1}{4} \leq \xi \leq 1$,

$$(2.16) \quad \mathcal{S}((p + \xi(q - x^*), t), r; (x^* + \xi(q - x^*), t), \varepsilon) \subset \Omega \times (0, T) \setminus \text{supp } v,$$

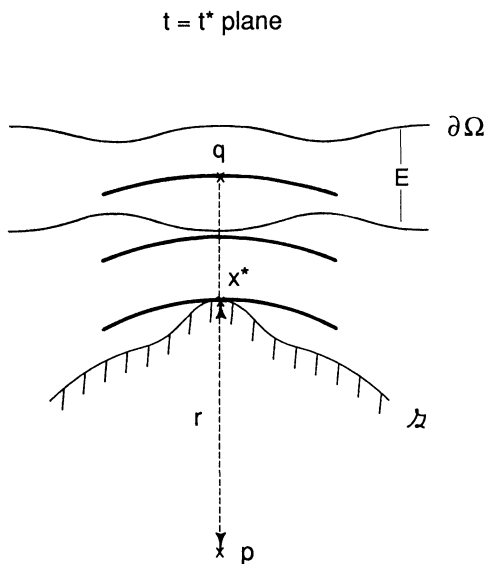


FIG. 1. $\mathcal{S}((p + \xi(q - x^*), t^*), r; (x^* + \xi(q - x^*), t^*), \varepsilon)$ for $\xi = 0, \frac{1}{2}$, and 1 when $n = 2$.

and, for all $|t - t^*| \leq \delta_2$ and $0 \leq \xi \leq 1$,

$$(2.17) \quad \begin{aligned} & \mathcal{S}((p + \xi(q - x^*), t), r; (x^* + \xi(q - x^*), t), \varepsilon) \\ & \setminus \mathcal{S}((p + \xi(q - x^*), t), r; (x^* + \xi(q - x^*), t), \tfrac{1}{2}\varepsilon) \\ & \subset \Omega \times (0, T) \setminus \text{supp } v. \end{aligned}$$

We now fix the above $t^*, x^*, p, q, r, \varepsilon$, and δ_2 to define

$$(2.18) \quad \begin{aligned} \Gamma(s) = & \bigcup_{|\xi| \leq 1/2} \mathcal{S}((p + (s + |\xi|)(q - x^*), t^* + \delta_2 \xi), r; \\ & (x^* + (s + |\xi|)(q - x^*), t^* + \delta_2 \xi), \varepsilon). \end{aligned}$$

Refer to Fig. 2 when $n = 2$. By (2.14), it is evident that

$$(2.19) \quad \Gamma(s) \subset \Omega \times (0, T), \quad \text{for } -\xi_0 \leq s \leq \tfrac{1}{2}.$$

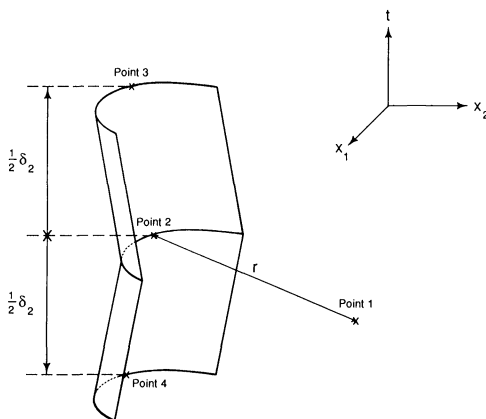


FIG. 2. $\Gamma(s)$ when $n = 2$; Point 1 = $(p + s(q - x^*), t^*)$; Point 2 = $(x^* + s(q - x^*), t^*)$; Point 3 = $(x^* + (s + \frac{1}{2})(q - x^*), t^* + \frac{1}{2}\delta_2)$; Point 4 = $(x^* + (s + \frac{1}{2})(q - x^*), t^* - \frac{1}{2}\delta_2)$.

LEMMA 2.3. *There is $0 \leq s^* < \frac{1}{4}$ such that*

$$(2.20) \quad \bigcup_{s^* < s \leq 1/2} \Gamma(s) \subset \Omega \times (0, T) \setminus \text{supp } v$$

and, for some $0 \leq |\xi^*| < \frac{1}{4}$,

$$(2.21) \quad \begin{aligned} & \mathcal{S}((p + (s^* + |\xi^*|)(q - x^*), t^* + \delta_2 \xi^*), r; \\ & (x^* + (s^* + |\xi^*|)(q - x^*), t^* + \delta_2 \xi^*), \tfrac{1}{2}\varepsilon) \\ & \cap \text{supp } v \text{ is not empty.} \end{aligned}$$

Proof. Let $s^* = \sup \{s \in [0, \frac{1}{2}]: \Gamma(s) \cap \text{supp } v \text{ is not empty}\}$. Then, $s^* < \frac{1}{4}$ by (2.16), and $\Gamma(s^*) \cap \text{supp } v$ is not empty. It is easy to see that the intersection of $\Gamma(s^*)$ and $\text{supp } v$ cannot occur on

$$\begin{aligned} & \mathcal{S}((p + (s^* + |\xi|)(q - x^*), t^* + \delta_2 \xi), r; (x^* + (s^* + |\xi|)(q - x^*), t^* + \delta_2 \xi), \varepsilon) \setminus \\ & \mathcal{S}((p + (s^* + |\xi|)(q - x^*), t^* + \delta_2 \xi), r; (x^* + (s^* + |\xi|)(q - x^*), t^* + \delta_2 \xi), \tfrac{1}{2}\varepsilon) \end{aligned}$$

for any $|\xi| \leq \frac{1}{2}$, according to (2.17). At the same time, (2.16) implies that the intersection cannot occur on

$$\mathcal{S}((p + (s^* + |\xi|)(q - x^*), t^* + \delta_2 \xi), r; (x^* + (s^* + |\xi|)(q - x^*), t^* + \delta_2 \xi), \varepsilon)$$

for $|\xi| \geq \frac{1}{4}$. Thus we conclude that there must be $0 \leq |\xi^*| < \frac{1}{4}$ such that (2.21) holds.

We now proceed to perform coordinate transformations. Let us fix s^* and ξ^* in (2.21), and choose a point

$$(2.22) \quad \begin{aligned} & (x_0, t_0) \in \text{supp } v \cap \mathcal{S}((p + (s^* + |\xi^*|)(q - x^*), t^* + \delta_2 \xi^*), r; \\ & (x^* + (s^* + |\xi^*|)(q - x^*), t^* + \delta_2 \xi^*), \tfrac{1}{2}\varepsilon). \end{aligned}$$

The first transformation. The set

$$\bigcup_{|\xi| \leq 1/2} \{(x^* + (s^* + |\xi|)(q - x^*), t^* + \delta_2 \xi)\}$$

is a central spine of $\Gamma(s^*)$. We first try to transform this wedge-shaped spine into a straight line segment. Let us set

$$(2.23) \quad \tau = t - t_0, \quad z = x - x_0 + \frac{1}{\delta_2} (x^* - q)(|t - t^*| - |t_0 - t^*|)$$

and

$$(2.24) \quad u(z, \tau) = v(x, t).$$

LEMMA 2.4. *There is a neighborhood Q_0 of $z = 0, \tau = 0$ in (z, τ) -space and positive numbers η_1, η_2 , and η_3 such that*

$$(2.25) \quad Q_0 \subset \text{domain of } u,$$

$$(2.26) \quad \bigcup_{\substack{|\xi| \leq \eta_1 \\ |\tau| \leq \eta_2}} \mathcal{S}(((1 + \xi)z^*, \tau), r; (\xi z^*, \tau), \eta_3) \subset Q_0,$$

$$(2.27) \quad u = 0 \text{ in } \bigcup_{\substack{-\eta_1 \leq \xi < 0 \\ |\tau| \leq \eta_2}} \mathcal{S}(((1 + \xi)z^*, \tau), r; (\xi z^*, \tau), \eta_3),$$

$$(2.28) \quad (0, 0) \in \text{supp } u,$$

where

$$(2.29) \quad z^* = p - x_0 + s^*(q - x^*) + \frac{1}{\delta_2} (q - x^*)|t_0 - t^*|.$$

Proof. Under the above transformation, the set $\Gamma(s)$ in (x, t) -space defined by (2.18) is transformed into the following set in (z, τ) -space:

$$(2.30) \quad \tilde{\Gamma}(s) = \bigcup_{|\xi| \leq 1/2} \mathcal{S}((z^* + (s - s^*)(q - x^*), \delta_2(\xi - \xi^*)), r; (z^* + (s - s^*)(q - x^*) + x^* - p, \delta_2(\xi - \xi^*)), \varepsilon).$$

We also note that the points $(p + (s^* + |\xi^*|)(q - x^*), t^* + \delta_2 \xi^*)$ and (x_0, t_0) in (x, t) -space are mapped into $(z^*, 0)$ and $(0, 0)$, respectively, in (z, τ) -space. Hence, (2.28) holds. By virtue of (2.22), it is apparent that $t_0 = t^* + \delta_2 \xi^*$ and $|p + (s^* + |\xi^*|)(q - x^*) - x_0| = r$. Thus, we have

$$(2.31) \quad |z^*| = r.$$

It also follows from (2.22) that

$$(2.32) \quad (0, 0) \in \text{supp } u \cap \mathcal{S}((z^*, 0), r; (z^* + x^* - p, 0), \tfrac{1}{2}\varepsilon)$$

in (z, τ) -space. Consequently, $(0, 0)$ is an interior point of

$$(2.33) \quad Q_0 \stackrel{\text{def}}{=} \text{the interior of } \bigcup \{ \tilde{\Gamma}(s) : |s - s^*| < \min \{ \tfrac{1}{4}, \xi_0, |x^* - p|/|q - x^*| \} \},$$

where ξ_0 is the positive number in (2.19). By (2.19), we find that

$$(2.34) \quad \tilde{\Gamma}(s) \subset \text{domain of } u$$

for each $-\xi_0 \leq s \leq \tfrac{1}{2}$, from which (2.25) follows.

Since $(0, 0)$ is an interior point of Q_0 and is the center point of the set $\mathcal{S}((z^*, 0), r; (0, 0), \eta)$ in (z, τ) -space for any $\eta \geq 0$, there are positive numbers η_1, η_2 , and η_3 such that (2.26) holds with

$$(2.35) \quad \eta_3 < \tfrac{1}{2}\varepsilon.$$

Next, we need to observe the following to complete the proof of Lemma 2.4.

LEMMA 2.5. *For each τ , it holds that*

$$(2.36) \quad \mathcal{S}(((1 + \xi)z^*, \tau), r; (\xi z^*, \tau), \eta_3) \cap \mathcal{S}((z^* + (s - s^*)(q - x^*), \tau), r; (z^* + (s - s^*)(q - x^*) + x^* - p, \tau), \varepsilon) \text{ is empty}$$

if $\xi < 0$ and $-|x^* - p|/|q - x^*| < s - s^* \leq 0$.

Proof. We first note that $\mathcal{S}(((1 + \xi)z^*, \tau), r; (\xi z^*, \tau), \eta_3)$ is a translation of $\mathcal{S}((z^*, \tau), r; (0, \tau), \eta_3)$ by the vector $(\xi z^*, 0)$ in R^{n+1} and that $\mathcal{S}((z^* + (s - s^*)(q - x^*), \tau), r; (z^* + (s - s^*)(q - x^*) + x^* - p, \tau), \varepsilon)$ is a translation of $\mathcal{S}((z^*, \tau), r; (z^* + x^* - p, \tau), \varepsilon)$ by the vector $((s - s^*)(q - x^*), 0)$ in R^{n+1} . We also recall that x^* belongs to the line segment between q and p . It follows from (2.32) and (2.35) that $\mathcal{S}((z^*, \tau), r; (0, \tau), \eta_3)$ is a subset of $\mathcal{S}((z^*, \tau), r; (z^* + x^* - p, \tau), \varepsilon)$. Hence (2.36) holds for $\xi < 0$ and $s = s^*$. Next, we suppose that $\xi < 0$ and $-|x^* - p|/|q - x^*| < s - s^* < 0$. Choose any point $(\rho_1, \tau) \in \mathcal{S}(((1 + \xi)z^*, \tau), r; (\xi z^*, \tau), \eta_3)$. Then, $|\rho_1 - (1 + \xi)z^*| = r$ and $|\rho_1 - \xi z^*| \leq \eta_3 < \tfrac{1}{2}\varepsilon < \tfrac{1}{2}r$. Hence, we have

$$(2.37) \quad \begin{aligned} |\rho_1 - z^*|^2 &= |\xi z^* + \rho_1 - (1 + \xi)z^*|^2 \\ &= \xi^2 r^2 + r^2 + 2\langle \xi z^*, \rho_1 - (1 + \xi)z^* \rangle \end{aligned}$$

and

$$(2.38) \quad \begin{aligned} \tfrac{1}{4}r^2 > |\rho_1 - \xi z^*|^2 &= |z^* + \rho_1 - (1 + \xi)z^*|^2 \\ &= r^2 + r^2 + 2\langle z^*, \rho_1 - (1 + \xi)z^* \rangle, \end{aligned}$$

where (2.31) has been used, and \langle, \rangle denotes the inner product in R^n . It follows from (2.38) that

$$\langle z^*, \rho_1 - (1 + \xi)z^* \rangle < 0,$$

which, together with (2.37), implies that

$$(2.39) \quad |\rho_1 - z^*|^2 > r^2.$$

On the other hand, we choose any

$$(\rho_2, \tau) \in \mathcal{S}((z^* + (s - s^*)(q - x^*), \tau), r; (z^* + (s - s^*)(q - x^*) + x^* - p, \tau), \varepsilon).$$

Then

$$|\rho_2 - (z^* + (s - s^*)(q - x^*))| = r$$

and

$$|\rho_2 - (z^* + (s - s^*)(q - x^*) + x^* - p)| \leq \varepsilon < r.$$

Thus, we see that

$$(2.40) \quad \begin{aligned} |\rho_2 - z^*|^2 &= |\rho_2 - (z^* + (s - s^*)(q - x^*)) + (s - s^*)(q - x^*)|^2 \\ &= r^2 + (s - s^*)^2 |q - x^*|^2 \\ &\quad + 2\langle \rho_2 - (z^* + (s - s^*)(q - x^*)), (s - s^*)(q - x^*) \rangle \end{aligned}$$

and

$$(2.41) \quad \begin{aligned} r^2 &> |\rho_2 - (z^* + (s - s^*)(q - x^*)) - (x^* - p)|^2 \\ &= r^2 + r^2 - 2\langle \rho_2 - (z^* + (s - s^*)(q - x^*)), x^* - p \rangle \end{aligned}$$

since $|x^* - p| = r$. It follows from (2.41) that

$$(2.42) \quad 2\langle \rho_2 - (z^* + (s - s^*)(q - x^*)), x^* - p \rangle > r^2.$$

Since x^* belongs to the line segment between q and p , and $-|x^* - p|/|q - x^*| < s - s^* < 0$, we can derive from (2.42) that

$$(2.43) \quad \begin{aligned} &2\langle \rho_2 - (z^* + (s - s^*)(q - x^*)), (s - s^*)(q - x^*) \rangle \\ &= (s - s^*)(|q - x^*|/|x^* - p|)2\langle \rho_2 - (z^* + (s - s^*)(q - x^*)), x^* - p \rangle \\ &< -|s - s^*||q - x^*|r. \end{aligned}$$

This combined with (2.40) yields

$$(2.44) \quad \begin{aligned} |\rho_2 - z^*|^2 &< r^2 + (s - s^*)^2 |q - x^*|^2 - |s - s^*||q - x^*|r \\ &< r^2 + |s - s^*||q - x^*|(r - r) = r^2 \end{aligned}$$

since $|s - s^*||q - x^*| < |x^* - p| = r$. By virtue of (2.39) and (2.44), (2.36) holds. This ends the proof of Lemma 2.5.

We proceed to prove (2.27). We can infer from (2.26), (2.33), and (2.36) that

$$(2.45) \quad \begin{aligned} &\bigcup_{\substack{-\eta_1 \leq \xi < 0 \\ |\tau| \leq \eta_2}} \mathcal{S}(((1 + \xi)z^*, \tau), r; (\xi z^*, \tau), \eta_3) \text{ is a subset of} \\ &\bigcup \{\tilde{\Gamma}(s) : |s - s^*| < \min(\tfrac{1}{4}, \xi_0, |x^* - p|/|q - x^*|), s > s^*\}. \end{aligned}$$

Meanwhile, we recall (2.20) to find that

$$(2.46) \quad u = 0 \quad \text{in} \quad \bigcup_{s^* < s \leq 1/2} \tilde{\Gamma}(s),$$

which, combined with (2.45), yields (2.27). The proof of Lemma 2.4 is now complete.

The second transformation. We now try to transform the central spine $\{(0, \tau): -\eta_2 \leq \tau \leq \eta_2\}$ into a wedge. The purpose is to satisfy (2.9). Set

$$(2.47) \quad \begin{aligned} s &= \tau, \\ y &= z + |\tau|z^*, \end{aligned}$$

and

$$(2.48) \quad w(y, s) = u(z, \tau).$$

Let Q be the image of Q_0 under this transformation. Then, Q is a neighborhood of $(0, 0)$ in (y, s) -space and

$$(2.49) \quad Q \subset \text{domain of } w.$$

It is also easy to see that $\mathcal{S}(((1 + \xi)z^*, \tau), r; (\xi z^*, \tau), \eta_3)$ in (z, τ) -space is transformed to $\mathcal{S}(((1 + \xi + |s|)y^*, s), r; ((\xi + |s|)y^*, s), \eta_3)$ in (y, s) -space, where we set

$$(2.50) \quad y^* = z^*.$$

Hence, it follows from (2.26)–(2.28) that

$$(2.51) \quad \bigcup_{\substack{|\xi| \leq \eta_1 \\ |s| \leq \eta_2}} \mathcal{S}(((1 + \xi + |s|)y^*, s), r; ((\xi + |s|)y^*, s), \eta_3) \subset Q,$$

$$(2.52) \quad w = 0 \quad \text{in} \quad \bigcup_{\substack{-\eta_1 \leq \xi < 0 \\ |s| \leq \eta_2}} \mathcal{S}(((1 + \xi + |s|)y^*, s), r; ((\xi + |s|)y^*, s), \eta_3),$$

$$(2.53) \quad (0, 0) \in \text{supp } w.$$

Next, we choose ε_1 , ε_2 , and ε_3 such that

$$(2.54) \quad \varepsilon_1 = \eta_3,$$

$$(2.55) \quad 0 < \varepsilon_2 < \frac{1}{2}\eta_2,$$

$$(2.56) \quad 2\varepsilon_2 + \varepsilon_3 < \eta_1,$$

$$(2.57) \quad 0 < \varepsilon_3 < \varepsilon_2.$$

Then, we have

$$(2.58) \quad |\xi - |s|| < \eta_1,$$

if $|\xi| \leq \varepsilon_3$ and $|s| \leq 2\varepsilon_2$. Consequently, if $|\tilde{\xi}| \leq \varepsilon_3$ and $|s| \leq 2\varepsilon_2$,

$$\begin{aligned} & \mathcal{S}(((1 + \tilde{\xi})y^*, s), r; (\tilde{\xi}y^*, s), \varepsilon_1) \\ &= \mathcal{S}(((1 + \tilde{\xi} - |s| + |s|)y^*, s), r; ((\tilde{\xi} - |s| + |s|)y^*, s), \varepsilon_1) \\ &\subset \bigcup_{|\xi| \leq \eta_1} \mathcal{S}(((1 + \xi + |s|)y^*, s), r; ((\xi + |s|)y^*, s), \eta_3). \end{aligned}$$

Thus, (2.8) holds. Next, it follows from (2.55)–(2.57) that

$$(2.59) \quad -\eta_1 < \xi - |s| < 0,$$

for $|\xi| \leq \varepsilon_3$ and $\varepsilon_2 \leq |s| \leq 2\varepsilon_2$, and that

$$(2.60) \quad -\eta_1 < \xi - |s| < 0,$$

for $-\varepsilon_3 \leq \xi < 0$ and $|s| \leq 2\varepsilon_2$. Now it is easy to see that

$$\begin{aligned}
 & \mathcal{S}(((1+\tilde{\xi})y^*, s), r; (\tilde{\xi}y^*, s), \varepsilon_1) \\
 (2.61) \quad &= \mathcal{S}(((1+\tilde{\xi}-|s|+|s|)y^*, s), r; (\tilde{\xi}-|s|+|s|)y^*, s), \eta_3) \\
 &\subset \bigcup_{\substack{-\eta_1 < \xi < 0 \\ |s| \leq \eta_2}} \mathcal{S}(((1+\xi+|s|)y^*, s), r; ((\xi+|s|)y^*, s), \eta_3)
 \end{aligned}$$

for $|\tilde{\xi}| \leq \varepsilon_3$ and $\varepsilon_2 \leq |s| \leq 2\varepsilon_2$, and that

$$\begin{aligned}
 & \mathcal{S}(((1+\tilde{\xi})y^*, s), r; (\tilde{\xi}y^*, s), \varepsilon_1) \\
 (2.62) \quad &\subset \bigcup_{\substack{-\eta_1 < \xi < 0 \\ |s| \leq \eta_2}} \mathcal{S}(((1+\xi+|s|)y^*, s), r; ((\xi+|s|)y^*, s), \eta_3)
 \end{aligned}$$

for $-\varepsilon_3 \leq \tilde{\xi} < 0$ and $|s| \leq 2\varepsilon_2$.

Now (2.9) and (2.10) follow from (2.52), (2.61), and (2.62). The proof of Proposition 2.2 is complete.

2.2. Proof of Theorem 2.1. We first present a special version of the Carleman estimates obtained by Isakov [3].

Let Q be a bounded open subset of R^{n+1} , and

$$(2.63) \quad \mathcal{P} = i \frac{\partial}{\partial s} + \Delta_y + i \sum_{j=1}^n \beta_j(s) \frac{\partial}{\partial y_j}, \quad i = \sqrt{-1},$$

where $\beta_j(s) \in L^\infty(Q)$, $j = 1, \dots, n$ are real-valued. Suppose that $\varphi \in C^\infty(\bar{Q})$ is a real-valued function such that for each $(y, s) \in \bar{Q}$,

$$(2.64) \quad \sum_{j=1}^n \left| \frac{\partial \varphi}{\partial y_j} \right| \neq 0,$$

$$(2.65) \quad \sum_{j,k=1}^n \zeta_j \bar{\zeta}_k \frac{\partial^2 \varphi}{\partial y_j \partial y_k} > 0,$$

for every nonzero complex vector $\zeta = (\zeta_1, \dots, \zeta_n) \in \mathbb{C}^n$.

Then, there is a positive constant C such that

$$(2.66) \quad \tau \int_Q \left(|u|^2 + \sum_{j=1}^n \left| \frac{\partial u}{\partial y_j} \right|^2 \right) \exp(2\tau\varphi) \, dy \, ds \leq C \int_Q |\mathcal{P}u|^2 \exp(2\tau\varphi) \, dy \, ds,$$

for all (complex-valued) $u \in C_0^\infty(Q)$ and all large $\tau > 0$. In fact, (2.66) holds for every u such that

$$(2.67) \quad \text{supp } u \text{ is a compact subset of } Q;$$

$$(2.68) \quad u, \quad \frac{\partial u}{\partial y_j} \in L^2(Q), \quad j = 1, \dots, n;$$

$$(2.69) \quad \mathcal{P}u \in L^2(Q).$$

For this, we argue as follows. Let $\mathcal{P}u = g \in L^2(Q)$ and $u^\varepsilon = u * \rho_\varepsilon$, where $\rho_\varepsilon \in C_0^\infty(R^{n+1})$ is the Friedrichs mollifier with sufficiently small ε so that $u^\varepsilon \in C_0^\infty(Q)$. Then, $u^\varepsilon(y, s)$ satisfies

$$\begin{aligned}
 (2.70) \quad & i \frac{\partial u^\varepsilon}{\partial s} + \Delta_y u^\varepsilon + i \sum_{j=1}^n \beta_j(s) \frac{\partial u^\varepsilon}{\partial y_j} \\
 &= g * \rho_\varepsilon + i \sum_{j=1}^n \beta_j(s) \frac{\partial u^\varepsilon}{\partial y_j} - \left(i \sum_{j=1}^n \beta_j \frac{\partial u}{\partial y_j} \right) * \rho_\varepsilon.
 \end{aligned}$$

We can now apply (2.66) to each u^ε and note that

$$(2.71) \quad g * \rho_\varepsilon \rightarrow g \quad \text{in } L^2(Q),$$

$$(2.72) \quad \sum_{j=1}^n \beta_j(s) \frac{\partial u^\varepsilon}{\partial y_j} - \left(\sum_{j=1}^n \beta_j \frac{\partial u}{\partial y_j} \right) * \rho_\varepsilon \rightarrow 0 \quad \text{in } L^2(Q)$$

since $\partial u / \partial y_j \in L^2(Q)$, $j = 1, \dots, n$. Thus the above claim has been justified.

We are now ready to prove Theorem 2.1. Assume that v is not almost everywhere zero in $\Omega \times (0, T)$. Then, by virtue of Proposition 2.2, there is a coordinate transformation (2.6), which satisfies (2.8)–(2.11). Meanwhile, $w(y, s)$ defined by (2.7) satisfies

$$(2.73) \quad \left(i \frac{\partial}{\partial s} + \Delta_y + i \sum_{j=1}^n \beta_j(s) \frac{\partial}{\partial y_j} \right) \cdot \left(-i \frac{\partial w}{\partial s} + \Delta_y w - i \sum_{j=1}^n \beta_j(s) \frac{\partial w}{\partial y_j} \right) = H(w),$$

where

$$\beta_j(s) = h'_j(s + t_0)$$

and

$$(2.74) \quad |H(w)| \leq M_2(y, s) \left(\left| \frac{\partial w}{\partial s} \right| + \sum_{|\alpha|=2} |\partial_y^\alpha w| \right) + M_1(y, s) \sum_{|\alpha| \leq 1} |\partial_y^\alpha w|,$$

with $M_2(y, s) \in L^\infty(Q)$ and $M_1(y, s) \in L^p(Q)$, $p > 2n$, which follow from (2.2)–(2.4). Here, $\partial_y^\alpha = (\partial / \partial y_1)^{\alpha_1} \cdots (\partial / \partial y_n)^{\alpha_n}$ and $h = (h_1, \dots, h_n)$ is the same as in (2.6).

Let

$$(2.75) \quad \lambda(y, s) = -i \frac{\partial w}{\partial s} + \Delta_y w - i \sum_{j=1}^n \beta_j(s) \frac{\partial w}{\partial y_j}.$$

Then, it holds that

$$(2.76) \quad \lambda, \frac{\partial \lambda}{\partial y_j} \in L^2(Q), \quad j = 1, \dots, n,$$

by the regularity of v , and that

$$(2.77) \quad \lambda = 0 \quad \text{in} \quad \bigcup_{\substack{|\xi| \leq \varepsilon_3 \\ \varepsilon_2 \leq |s| \leq 2\varepsilon_2}} \mathcal{S}(((1 + \xi)y^*, s), r; (\xi y^*, s), \varepsilon_1),$$

$$(2.78) \quad \lambda = 0 \quad \text{in} \quad \bigcup_{\substack{-\varepsilon_3 \leq \xi < 0 \\ |s| \leq 2\varepsilon_2}} \mathcal{S}(((1 + \xi)y^*, s), r; (\xi y^*, s), \varepsilon_1),$$

according to (2.9) and (2.10). We choose

$$(2.79) \quad 0 < \varepsilon_4 < \frac{1}{2} \min(\varepsilon_1, \varepsilon_3 |y^*|, r)$$

and $\psi(y) \in C_0^\infty(\mathbb{R}^n)$ such that

$$(2.80) \quad \begin{aligned} \psi &= 1 && \text{for } |y| \leq \frac{1}{2}\varepsilon_4, \\ \psi &= 0 && \text{for } |y| > \frac{2}{3}\varepsilon_4. \end{aligned}$$

We then set

$$(2.81) \quad Q_1 = \{(y, s) : |y| < \varepsilon_4, |s| < 2\varepsilon_2\}$$

and

$$(2.82) \quad \sigma(y, s) = \psi(y)\lambda(y, s).$$

By virtue of (2.8) and (2.9), it is easy to see that $Q_1 \subset Q$

$$(2.83) \quad \text{supp } \sigma \subset Q_1,$$

$$(2.84) \quad \sigma, \quad \frac{\partial \sigma}{\partial y_j} \in L^2(Q_1), \quad j = 1, \dots, n,$$

and σ satisfies

$$(2.85) \quad \begin{aligned} i \frac{\partial \sigma}{\partial s} + \Delta_y \sigma + \sum_{j=1}^n i \beta_j(s) \frac{\partial \sigma}{\partial y_j} &= \psi(y) H(w) + 2 \sum_{j=1}^n \frac{\partial \psi}{\partial y_j} \frac{\partial \lambda}{\partial y_j} \\ &\quad + \lambda \Delta_y \psi + i \sum_{j=1}^n \beta_j(s) \lambda \frac{\partial \psi}{\partial y_j}. \end{aligned}$$

Next, we define

$$(2.86) \quad \varphi(y) = |y - (2 + \varepsilon_3)y^*|^2.$$

Then, φ satisfies (2.64) and (2.65) on \bar{Q}_1 . Hence, we can apply (2.66) to σ as follows:

$$(2.87) \quad \begin{aligned} &\tau \int_{Q_1} \left(|\sigma|^2 + \sum_{j=1}^n \left| \frac{\partial \sigma}{\partial y_j} \right|^2 \right) \exp(2\tau\varphi) \, dy \, ds \\ &\leq C \int_{Q_1} \left| \psi(y) H(w) + 2 \sum_{j=1}^n \frac{\partial \psi}{\partial y_j} \frac{\partial \lambda}{\partial y_j} + \lambda \Delta_y \psi \right. \\ &\quad \left. + i \sum_{j=1}^n \beta_j(s) \lambda \frac{\partial \psi}{\partial y_j} \right|^2 \exp(2\tau\varphi) \, dy \, ds \\ &\leq C \int_{Q_1} |H(w)|^2 \exp(2\tau\varphi) \, dy \, ds \\ &\quad + C \int_{Q_1 \setminus Q_2} \left(\sum_{|\alpha| \leq 1} \left| \partial_y^\alpha \frac{\partial w}{\partial s} \right|^2 + \sum_{|\alpha| \leq 3} |\partial_y^\alpha w|^2 \right) \exp(2\tau\varphi) \, dy \, ds, \end{aligned}$$

where

$$(2.88) \quad Q_2 = \{(y, s) : |y| < \frac{1}{2}\varepsilon_4, |s| < 2\varepsilon_2\}.$$

Using the same $\psi(y)$ as above, we set

$$(2.89) \quad \zeta(y, s) = \psi(y)w(y, s).$$

Then, $\zeta(y, s)$ satisfies

$$(2.90) \quad \text{supp } \zeta \subset Q_1,$$

$$(2.91) \quad \partial_y^\alpha \zeta \in L^2(Q_1), \quad |\alpha| \leq 3,$$

$$(2.92) \quad \begin{aligned} -i \frac{\partial \zeta}{\partial s} + \Delta_y \zeta - i \sum_{j=1}^n \beta_j(s) \frac{\partial \zeta}{\partial y_j} &= \sigma(y, s) + 2 \sum_{j=1}^n \frac{\partial \psi}{\partial y_j} \frac{\partial w}{\partial y_j} \\ &\quad + w \Delta_y \psi - i \sum_{j=1}^n \beta_j(s) w \frac{\partial \psi}{\partial y_j}, \end{aligned}$$

$$\begin{aligned}
 (2.93) \quad & -i \frac{\partial}{\partial s} \frac{\partial \zeta}{\partial y_m} + \Delta_y \frac{\partial \zeta}{\partial y_m} - i \sum_{j=1}^n \beta_j(s) \frac{\partial}{\partial y_j} \frac{\partial \zeta}{\partial y_m} \\
 & = \frac{\partial}{\partial y_m} \left\{ \sigma(y, s) + 2 \sum_{j=1}^n \frac{\partial \psi}{\partial y_j} \frac{\partial w}{\partial y_j} + w \Delta_y \psi - i \sum_{j=1}^n \beta_j(s) w \frac{\partial \psi}{\partial y_j} \right\}, \\
 & \qquad \qquad \qquad m = 1, \dots, n.
 \end{aligned}$$

By applying (2.66) to ζ and $\partial \zeta / \partial y_m$, $m = 1, \dots, n$, with the aid of (2.87), we obtain

$$\begin{aligned}
 (2.94) \quad & \tau \int_{Q_1} \sum_{|\alpha| \leq 2} |\partial_y^\alpha \zeta|^2 \exp(2\tau\varphi) \, dy \, ds \\
 & \leq \frac{C}{\tau} \int_{Q_1} |H(w)|^2 \exp(2\tau\varphi) \, dy \, ds \\
 & \quad + C \int_{Q_1 \setminus Q_2} \left(\sum_{|\alpha| \leq 1} \left| \partial_y^\alpha \frac{\partial w}{\partial s} \right|^2 + \sum_{|\alpha| \leq 3} |\partial_y^\alpha w|^2 \right) \exp(2\tau\varphi) \, dy \, ds
 \end{aligned}$$

for all large $\tau > 0$.

Next, we use (2.92) to obtain

$$\begin{aligned}
 (2.95) \quad & \int_{Q_1} \left| \frac{\partial \zeta}{\partial s} \right|^2 \exp(2\tau\varphi) \, dy \, ds \leq C \int_{Q_1} \sum_{|\alpha| \leq 2} |\partial_y^\alpha \zeta|^2 \exp(2\tau\varphi) \, dy \, ds \\
 & \quad + C \int_{Q_1} |\sigma(y, s)|^2 \exp(2\tau\varphi) \, dy \, ds \\
 & \quad + C \int_{Q_1 \setminus Q_2} \sum_{|\alpha| \leq 1} |\partial_y^\alpha w|^2 \exp(2\tau\varphi) \, dy \, ds.
 \end{aligned}$$

Combining (2.87), (2.94), and (2.95), we have

$$\begin{aligned}
 (2.96) \quad & \int_{Q_1} \left(\left| \frac{\partial \zeta}{\partial s} \right|^2 + \sum_{|\alpha| \leq 2} |\partial_y^\alpha \zeta|^2 \right) \exp(2\tau\varphi) \, dy \, ds \\
 & \leq \frac{C}{\tau} \int_{Q_1} |H(w)|^2 \exp(2\tau\varphi) \, dy \, ds \\
 & \quad + \frac{C}{\tau} \int_{Q_1 \setminus Q_2} \left(\sum_{|\alpha| \leq 1} \left| \partial_y^\alpha \frac{\partial w}{\partial s} \right|^2 + \sum_{|\alpha| \leq 3} |\partial_y^\alpha w|^2 \right) \exp(2\tau\varphi) \, dy \, ds \\
 & \quad + C \int_{Q_1 \setminus Q_2} \sum_{|\alpha| \leq 1} |\partial_y^\alpha w|^2 \exp(2\tau\varphi) \, dy \, ds,
 \end{aligned}$$

for all large $\tau > 0$.

Next we recall (2.74) to find that

$$\begin{aligned}
 (2.97) \quad & \int_{Q_1} |H(w)|^2 \exp(2\tau\varphi) \, dy \, ds \leq C \int_{Q_1} \left(\left| \frac{\partial w}{\partial s} \right|^2 + \sum_{|\alpha| \leq 2} |\partial_y^\alpha w|^2 \right) \exp(2\tau\varphi) \, dy \, ds \\
 & \quad + C \int_{Q_1} M_1(y, s)^2 \sum_{|\alpha| \leq 1} |\partial_y^\alpha w|^2 \exp(2\tau\varphi) \, dy \, ds,
 \end{aligned}$$

for all τ . Since $M_1(y, s) \in L^{2p}(Q_1)$, $p > n$, it is easy to see that

$$(2.98) \quad \int_{Q_1} M_1(y, s)^2 |\partial_y^\alpha w|^2 \exp(2\tau\varphi) \, dy \, ds \leq C \left(\int_{Q_1} (|\partial_y^\alpha w| \exp(\tau\varphi))^{2q} \, dy \, ds \right)^{1/q}$$

for all τ and $|\alpha| \leq 1$, where $q = p/(p-1) < n/(n-1)$.

Since $w = 0$ in $Q_1 \cap \{(y, s): |s| \geq \varepsilon_2\}$, there is a bounded open set $Q_1^* \subset Q_1$ such that ∂Q_1^* is C^1 and $w = 0$ in $Q_1 \setminus Q_1^*$. Hence, we can apply a special version of the Gagliardo–Nirenberg–Sobolev inequality to find that

$$(2.99) \quad \begin{aligned} & \|(\partial_y^\alpha w) \exp(\tau\varphi)\|_{L^{2q}(Q_1^*)} \leq C \|(\partial_y^\alpha w) \exp(\tau\varphi)\|_{L^2(Q_1^*)}^{\frac{1}{\nu}} \\ & \cdot \left(\sum_{j=1}^n \|(\partial_y^\alpha w) \tau \frac{\partial \varphi}{\partial y_j} \exp(\tau\varphi)\|_{L^2(Q_1^*)} \right. \\ & \quad \left. + \sum_{|\beta| \geq 2} \|(\partial_y^\beta w) \exp(\tau\varphi)\|_{L^2(Q_1^*)} \right)^\nu, \end{aligned}$$

for all τ and $|\alpha| = 1$, and

$$(2.100) \quad \begin{aligned} & \|w \exp(\tau\varphi)\|_{L^{2q}(Q_1^*)} \leq C \|w \exp(\tau\varphi)\|_{L^2(Q_1^*)}^{\frac{1}{\nu}} \\ & \cdot \left(\sum_{j=1}^n \left\| w \tau \frac{\partial \varphi}{\partial y_j} \exp(\tau\varphi) \right\|_{L^2(Q_1^*)} \right. \\ & \quad \left. + \sum_{|\beta| \geq 1} \|(\partial_y^\beta w) \exp(\tau\varphi)\|_{L^2(Q_1^*)} \right)^\nu, \end{aligned}$$

for all τ , where $\nu = n(q-1)/2q < \frac{1}{2}$. Combining (2.97)–(2.100), we obtain

$$(2.101) \quad \begin{aligned} & \int_{Q_1} |H(w)|^2 \exp(2\tau\varphi) \, dy \, ds \leq C \int_{Q_1} \left(\left| \frac{\partial w}{\partial s} \right|^2 + \sum_{|\alpha| \geq 2} |\partial_y^\alpha w|^2 \right) \exp(2\tau\varphi) \, dy \, ds \\ & \quad + C\tau^{2\nu} \int_{Q_1} \sum_{|\alpha| \geq 2} |\partial_y^\alpha w|^2 \exp(2\tau\varphi) \, dy \, ds, \end{aligned}$$

for all large $\tau > 0$.

By taking τ sufficiently large, we derive from (2.80), (2.89), (2.96), and (2.101) that

$$(2.102) \quad \begin{aligned} & \int_{Q_2} \left(\left| \frac{\partial w}{\partial s} \right|^2 + \sum_{|\alpha| \geq 2} |\partial_y^\alpha w|^2 \right) \exp(2\tau\varphi) \, dy \, ds \\ & \leq C \int_{Q_1 \setminus Q_2} \left(\sum_{|\alpha| \leq 1} \left| \partial_y^\alpha \frac{\partial w}{\partial s} \right|^2 + \sum_{|\alpha| \geq 3} |\partial_y^\alpha w|^2 \right) \exp(2\tau\varphi) \, dy \, ds. \end{aligned}$$

Now we recall that $w = 0$ in

$$\bigcup_{\substack{-\varepsilon_3 \leq \xi < 0 \\ |s| \geq 2\varepsilon_2}} \mathcal{S}(((1+\xi)y^*, s), r; (\xi y^*, s), \varepsilon_1).$$

Set

$$(2.103) \quad \mathcal{T} = \bigcup_{\substack{0 \leq \xi \leq \varepsilon_3 \\ |s| \geq 2\varepsilon_2}} \mathcal{S}(((1+\xi)y^*, s), r; (\xi y^*, s), \varepsilon_1)$$

and

$$(2.104) \quad Q_3 = \{(y, s): |y| < \varepsilon_5, |s| < 2\varepsilon_2\},$$

where $0 < \varepsilon_5 < \frac{1}{2}\varepsilon_4$. It then follows from (2.102) that

$$(2.105) \quad \begin{aligned} & \int_{\mathcal{T} \cap Q_3} \left(\left| \frac{\partial w}{\partial s} \right|^2 + \sum_{|\alpha| \geq 2} |\partial_y^\alpha w|^2 \right) \exp(2\tau\varphi) \, dy \, ds \\ & \leq C \int_{\mathcal{T} \cap (Q_1 \setminus Q_2)} \left(\sum_{|\alpha| \leq 1} \left| \partial_y^\alpha \frac{\partial w}{\partial s} \right|^2 + \sum_{|\alpha| \geq 3} |\partial_y^\alpha w|^2 \right) \exp(2\tau\varphi) \, dy \, ds. \end{aligned}$$

We can choose ε_5 so small that

$$(2.106) \quad \varphi \leq \kappa_1 \quad \text{on } \mathcal{T} \cap (Q_1 \setminus Q_2)$$

and

$$(2.107) \quad \varphi \geq \kappa_2 \quad \text{on } \mathcal{T} \cap Q_3$$

for some positive constants $\kappa_1 < \kappa_2$ (see Fig. 3). Then, (2.105) yields

$$(2.108) \quad \begin{aligned} & \int_{\mathcal{T} \cap Q_3} \left(\left| \frac{\partial w}{\partial s} \right|^2 + \sum_{|\alpha| \leq 2} |\partial_y^\alpha w|^2 \right) \exp(2\tau\kappa_2) \, dy \, ds \\ & \leq C \int_{\mathcal{T} \cap (Q_1 \setminus Q_2)} \left(\sum_{|\alpha| \leq 1} \left| \partial_y^\alpha \frac{\partial w}{\partial s} \right|^2 + \sum_{|\alpha| \leq 3} |\partial_y^\alpha w|^2 \right) \exp(2\tau\kappa_1) \, dy \, ds. \end{aligned}$$

By passing $\tau \rightarrow \infty$, we find that $w = 0$ in Q_3 , which contradicts the assumption that $(0, 0)$ belongs to supp w . This completes the proof of Theorem 2.1.

3. Proof of Proposition 1.6. According to Theorem 2.1, it is enough to show that $v \in L^2(0, T; H^3(\Omega))$ and $v_t \in L^2(0, T; H^1(\Omega))$. Throughout this section, we assume (1.2) and (1.3), and M denotes positive constants independent of given functions.

LEMMA 3.1. *For given $g(x, t) \in L^1(0, T; L^2(\Omega))$, there is a unique function Φ in $C([0, T]; H_0^2(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ that satisfies*

$$(3.1) \quad \Phi_{tt} + \Delta^2 \Phi + \sum_{|\alpha| \leq 2} (-1)^{|\alpha|} \partial_x^\alpha (b_\alpha(x, t)\Phi) - \partial_t(d(x, t)\Phi) = g(x, t) \quad \text{in } \Omega \times (0, T),$$

$$(3.2) \quad \Phi(x, 0) = \Phi_t(x, 0) = 0 \quad \text{in } \Omega.$$

Furthermore, it holds that

$$(3.3) \quad \|\Phi\|_{C([0, T]; H_0^2(\Omega))} + \|\Phi_t\|_{C([0, T]; L^2(\Omega))} \leq M \|g\|_{L^1(0, T; L^2(\Omega))},$$

$$(3.4) \quad \|\Phi\|_{C([0, T]; L^2(\Omega))} + \|\Phi_t\|_{C([0, T]; H^{-2}(\Omega))} \leq M \|g\|_{L^1(0, T; H^{-2}(\Omega))},$$

$$(3.5) \quad \|\Phi\|_{C([0, T]; H_0^1(\Omega))} + \|\Phi_t\|_{C([0, T]; H^{-1}(\Omega))} \leq M \|g\|_{L^1(0, T; H^{-1}(\Omega))}.$$

This is a well-known fact; see [10] and [13] for the relevant technical details.

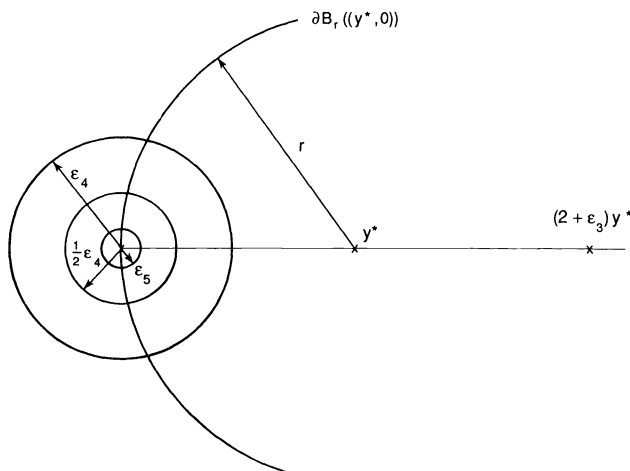


FIG. 3. Cross sections of Q_1 , Q_2 , and Q_3 at $s = 0$.

LEMMA 3.2. If $\Phi \in C([0, T]; H_0^2(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ satisfies (3.1) with $g \equiv 0$ in $\Omega \times (0, T)$, then

$$(3.6) \quad \|\Phi\|_{C([0, T]; H_0^2(\Omega))}^2 + \|\Phi_t\|_{C([0, T]; L^2(\Omega))}^2 \leq M \left(\int_0^T \int_E \Phi_t^2 dx dt + \int_0^T \int_\Omega \Phi^2 dx dt \right),$$

where E is a neighborhood of $\partial\Omega$, as in the previous section.

This can be proved by using the same argument as in the proof of Proposition 1.4. The unique continuation property is not an issue on account of the last integral in (3.6).

LEMMA 3.3. Suppose that $w \in C([0, T]; H_0^2(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ satisfies (1.28) in $\Omega \times (0, T)$. Then, we have

$$(3.7) \quad \begin{aligned} & \|w\|_{C([0, T]; H_0^2(\Omega))}^2 + \|w_t\|_{C([0, T]; L^2(\Omega))}^2 \\ & \leq M \left(\int_0^T \int_E w_t^2 dx dt + \|w_t\|_{L^2(0, T; H^{-2}(\Omega))}^2 \right). \end{aligned}$$

Proof. If $w \in C([0, T]; H_0^2(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ satisfies (1.28) in $\Omega \times (0, T)$ and $w_t = 0$ in $\Omega \times (0, T)$, then $w = w(x) \in H_0^2(\Omega)$ satisfies

$$(3.8) \quad \Delta^2 w + \sum_{|\alpha| \geq 2} (-1)^{|\alpha|} \partial_x^\alpha (b_\alpha(x, 0)w) + Kw = 0 \quad \text{in } \Omega,$$

and, consequently, $w = 0$ in $\Omega \times (0, T)$. By virtue of this observation, we can repeat the same argument as in the proof of Proposition 1.4 to prove (3.7).

LEMMA 3.4. If $\Phi \in C([0, T]; H_0^2(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ satisfies (3.1) with $g \equiv 0$ in $\Omega \times (0, T)$, then

$$(3.9) \quad \begin{aligned} & \|\Phi\|_{C([0, T]; L^2(\Omega))}^2 + \|\Phi_t\|_{C([0, T]; H^{-2}(\Omega))}^2 \\ & \leq M \left(\int_0^T \int_E \Phi^2 dx dt + \|\Phi\|_{L^2(0, T; H^{-2}(\Omega))}^2 \right). \end{aligned}$$

Proof. As in the proof of Lemma 1.3, we set

$$(3.10) \quad w(x, t) = \int_0^t \Phi(x, t) dt + \chi(x),$$

where χ is a unique solution in $H_0^2(\Omega)$ of

$$(3.11) \quad \Delta^2 \chi + \sum_{|\alpha| \geq 2} (-1)^{|\alpha|} \partial_x^\alpha (b_\alpha(x, 0)\chi) + K\chi = -\Phi_t(x, 0) + d(x, 0)\Phi(x, 0).$$

Then, w satisfies (1.28) in $\Omega \times (0, T)$. Now (3.9) follows from (3.7).

The following proposition completes the proof of Proposition 1.6.

PROPOSITION 3.5. If $v \in C([0, T]; L^2(\Omega)) \cap C^1([0, T]; H^{-2}(\Omega))$ satisfies (1.13) in $\Omega \times (0, T)$ and $v = 0$ in $E \times (0, T)$, then $v \in C([0, T]; H^3(\Omega)) \cap C^1([0, T]; H^1(\Omega))$.

Proof. Let $v^\varepsilon = v * \rho_\varepsilon$, where $\rho_\varepsilon \in C_0^\infty(R^n)$ is the Friedrichs mollifier and the convolution is taken with respect to the space variables. We choose a neighborhood E' of $\partial\Omega$ and $\varepsilon_0 > 0$ such that $E' \subset E$ and $v^\varepsilon = 0$ in $E' \times (0, T)$ for each $0 < \varepsilon < \varepsilon_0$. From now on, we restrict to such ε . Obviously, $v^\varepsilon, v_{tt}^\varepsilon \in L^\infty(0, T; C^\infty(\bar{\Omega}))$, and v^ε satisfies

$$(3.12) \quad v_{tt}^\varepsilon + \Delta^2 v^\varepsilon + \sum_{|\alpha| \geq 2} (-1)^{|\alpha|} \partial_x^\alpha (b_\alpha(x, t)v^\varepsilon) - \partial_t(d(x, t)v^\varepsilon) = G^\varepsilon(x, t),$$

where

$$(3.13) \quad G^\varepsilon = \sum_{|\alpha| \geq 2} (-1)^{|\alpha|} \partial_x^\alpha (b_\alpha v^\varepsilon) - \sum_{|\alpha| \geq 2} (-1)^{|\alpha|} \partial_x^\alpha (b_\alpha v) * \rho_\varepsilon - \partial_t(dv^\varepsilon) + \partial_t(dv) * \rho_\varepsilon.$$

Then, $G^\varepsilon \in L^\infty(\Omega \times (0, T))$. By virtue of (1.2) and (1.3), we can use Friedrichs's lemma to find that, for every $\varepsilon > 0$,

$$(3.14) \quad \|\partial_x^\alpha(b_\alpha v^\varepsilon) - \partial_x^\alpha(b_\alpha v) * \rho_\varepsilon\|_{L^\infty(0, T; H^{-1}(\Omega))} \leq M \quad \text{for } |\alpha| = 2,$$

$$(3.15) \quad \|dv_t^\varepsilon - (dv_t) * \rho_\varepsilon\|_{L^\infty(0, T; H^{-1}(\Omega))} \leq M.$$

It is easy to see that the remaining terms of G^ε are also bounded in $L^\infty(0, T; H^{-1}(\Omega))$ uniformly in ε . Hence, we have

$$(3.16) \quad \|G^\varepsilon\|_{L^\infty(0, T; H^{-1}(\Omega))} \leq M \quad \text{for all } \varepsilon.$$

Next, we fix any $j = 1, \dots, n$, and write

$$(3.17) \quad w^\varepsilon = \partial_j v^\varepsilon,$$

where $\partial_j = \partial/\partial x_j$. Then, w^ε satisfies

$$(3.18) \quad w_{tt}^\varepsilon + \Delta^2 w^\varepsilon + \sum_{|\alpha| \leq 2} (-1)^{|\alpha|} \partial_x^\alpha(b_\alpha w^\varepsilon) - \partial_t(dw^\varepsilon) = H^\varepsilon \quad \text{in } \Omega \times (0, T),$$

where

$$(3.19) \quad H^\varepsilon = \partial_j G^\varepsilon - \sum_{|\alpha| \leq 2} (-1)^{|\alpha|} \partial_x^\alpha((\partial_j b_\alpha) v^\varepsilon) + \partial_t((\partial_j d) v^\varepsilon).$$

Again by (1.2) and (1.3), $H^\varepsilon \in L^\infty(\Omega \times (0, T))$ and

$$(3.20) \quad \|H^\varepsilon\|_{L^\infty(0, T; H^{-2}(\Omega))} \leq M \quad \text{for all } \varepsilon.$$

Next, let Φ^ε be a unique solution in $C([0, T]; H_0^2(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ of

$$(3.21) \quad \Phi_{tt}^\varepsilon + \Delta^2 \Phi^\varepsilon + \sum_{|\alpha| \leq 2} (-1)^{|\alpha|} \partial_x^\alpha(b_\alpha \Phi^\varepsilon) - \partial_t(d\Phi^\varepsilon) = H^\varepsilon \quad \text{in } \Omega \times (0, T),$$

$$(3.22) \quad \Phi^\varepsilon(x, 0) = \Phi_t^\varepsilon(x, 0) = 0 \quad \text{in } \Omega.$$

By Lemma 3.1, we have

$$(3.23) \quad \|\Phi^\varepsilon\|_{C([0, T]; L^2(\Omega))} + \|\Phi_t^\varepsilon\|_{C([0, T]; H^{-2}(\Omega))} \leq M,$$

for all ε . Let Θ^ε be a unique solution in $C([0, T]; H_0^2(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ of

$$(3.24) \quad \Theta_{tt}^\varepsilon + \Delta^2 \Theta^\varepsilon + \sum_{|\alpha| \leq 2} (-1)^{|\alpha|} \partial_x^\alpha(b_\alpha \Theta^\varepsilon) - \partial_t(d\Theta^\varepsilon) = 0 \quad \text{in } \Omega \times (0, T),$$

$$(3.25) \quad \Theta^\varepsilon(x, 0) = w^\varepsilon(x, 0), \quad \Theta_t^\varepsilon(x, 0) = w_t^\varepsilon(x, 0) \quad \text{in } \Omega.$$

By the uniqueness of solution, we have

$$(3.26) \quad w^\varepsilon = \Theta^\varepsilon + \Phi^\varepsilon \quad \text{in } \Omega \times (0, T),$$

and thus

$$(3.27) \quad \Theta^\varepsilon = -\Phi^\varepsilon \quad \text{in } E' \times (0, T).$$

It follows from Lemma 3.4, (3.23), (3.26), (3.27), and

$$(3.28) \quad \|v^\varepsilon\|_{C([0, T]; L^2(\Omega))} \leq M \quad \text{for all } \varepsilon$$

that

$$(3.29) \quad \begin{aligned} & \|\Theta^\varepsilon\|_{C([0, T]; L^2(\Omega))}^2 + \|\Theta_t^\varepsilon\|_{C([0, T]; H^{-2}(\Omega))}^2 \\ & \leq M \left(\int_0^T \int_{E'} (\Phi^\varepsilon)^2 dx dt + \|w^\varepsilon\|_{L^2(0, T; H^{-2}(\Omega))}^2 + \|\Phi^\varepsilon\|_{L^2(0, T; H^{-2}(\Omega))}^2 \right) \\ & \leq M \quad \text{for all } \varepsilon. \end{aligned}$$

We can now conclude from (3.23), (3.26), and (3.29) that

$$(3.30) \quad \|w^\varepsilon\|_{C([0,T];L^2(\Omega))} + \|w_t^\varepsilon\|_{C([0,T];H^{-2}(\Omega))} \leq M,$$

for all ε , and that

$$(3.31) \quad \|v^\varepsilon\|_{C([0,T];H_0^1(\Omega))} + \|v_t^\varepsilon\|_{C([0,T];H^{-1}(\Omega))} \leq M,$$

for all ε . This improves the estimate of G^ε . Again by means of Friedrichs's lemma, it follows from (3.31) that

$$(3.32) \quad \|\partial_x^\alpha(b_\alpha v^\varepsilon) - \partial_x^\alpha(b_\alpha v) * \rho_\varepsilon\|_{L^\infty(0,T;L^2(\Omega))} \leq M \quad \text{for } |\alpha| = 2,$$

$$(3.33) \quad \|dv_t^\varepsilon - (dv_t) * \rho_\varepsilon\|_{L^\infty(0,T;L^2(\Omega))} \leq M,$$

for all ε . It is apparent that the remaining terms of G^ε are also bounded in $L^\infty(0, T; L^2(\Omega))$ uniformly in ε . Hence, we have

$$(3.34) \quad \|G^\varepsilon\|_{L^\infty(0,T;L^2(\Omega))} \leq M \quad \text{for all } \varepsilon.$$

Next, we denote by Ψ^ε a unique solution in $C([0, T]; H_0^2(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ of

$$(3.35) \quad \Psi_{tt}^\varepsilon + \Delta^2 \Psi^\varepsilon + \sum_{|\alpha| \leq 2} (-1)^{|\alpha|} \partial_x^\alpha(b_\alpha \Psi^\varepsilon) - \partial_t(d\Psi^\varepsilon) = G^\varepsilon \quad \text{in } \Omega \times (0, T),$$

$$(3.36) \quad \Psi^\varepsilon(x, 0) = \Psi_t^\varepsilon(x, 0) = 0 \quad \text{in } \Omega.$$

By virtue of (3.3) and (3.34), it follows that

$$(3.37) \quad \|\Psi^\varepsilon\|_{C([0,T];H_0^2(\Omega))} + \|\Psi_t^\varepsilon\|_{C([0,T];L^2(\Omega))} \leq M,$$

for all ε . We then define Ξ^ε to be a unique solution in $C([0, T]; H_0^2(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ of

$$(3.38) \quad \Xi_{tt}^\varepsilon + \Delta^2 \Xi^\varepsilon + \sum_{|\alpha| \leq 2} (-1)^{|\alpha|} \partial_x^\alpha(b_\alpha \Xi^\varepsilon) - \partial_t(d\Xi^\varepsilon) = 0 \quad \text{in } \Omega \times (0, T),$$

$$(3.39) \quad \Xi^\varepsilon(x, 0) = v^\varepsilon(x, 0), \quad \Xi_t^\varepsilon(x, 0) = v_t^\varepsilon(x, 0),$$

so that

$$(3.40) \quad v^\varepsilon = \Psi^\varepsilon + \Xi^\varepsilon \quad \text{in } \Omega \times (0, T).$$

It follows from Lemma 3.2, (3.37) and (3.40) that

$$(3.41) \quad \|\Xi^\varepsilon\|_{C([0,T];H_0^2(\Omega))} + \|\Xi_t^\varepsilon\|_{C([0,T];L^2(\Omega))} \leq M,$$

for all ε , and, consequently,

$$(3.42) \quad \|v^\varepsilon\|_{C([0,T];H_0^2(\Omega))} + \|v_t^\varepsilon\|_{C([0,T];L^2(\Omega))} \leq M$$

for all ε . By passing $\varepsilon \rightarrow 0$, we conclude that

$$(3.43) \quad v \in L^\infty(0, T; H_0^2(\Omega)), \quad v_t \in L^\infty(0, T; L^2(\Omega)),$$

which, in fact, yields

$$(3.44) \quad v \in C([0, T]; H_0^2(\Omega)), \quad v_t \in C([0, T]; L^2(\Omega)).$$

Finally, we differentiate (1.13) by ∂_j , $j = 1, \dots, n$, and arrive at

$$(3.45) \quad v \in C([0, T]; H^3(\Omega)), \quad v_t \in C([0, T]; H^1(\Omega))$$

through a similar argument. This ends the proof of Proposition 3.5.

Acknowledgments. The author acknowledges a very helpful conversation with Professor E. Zuazua in Vorau, Austria. He suggested various improvements of the earlier version of this paper. Anonymous referees also made significant contributions in revising the paper. The author thanks all of them.

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, Masson, Paris, 1988.
- [2] A. HARAUX, *Séries lacunaires et contrôle semi-interne des vibrations d'une plaque rectangulaire*, J. Math. Pures Appl., 68 (1989), pp. 457–465.
- [3] V. M. ISAKOV, *On the uniqueness of the solution of the Cauchy problem*, Soviet Math. Dokl., 22 (1980), pp. 639–642.
- [4] S. JAFFARD, *Contrôle interne exact des vibrations d'une plaque carrée*, C. R. Acad. Sci. Paris, 307 (1988), pp. 759–762.
- [5] ———, *Contrôle interne des vibrations d'une plaque rectangulaire*, Portugal Math., 47 (1990), pp. 423–429.
- [6] H. KHALGUI-OUNAIES, *Unicité de problème de Cauchy pour les opérateurs quasi-homogènes*, Ann. Scuola Norm. Sup. Pisa, 15 (1988), pp. 567–582.
- [7] J. KIM, *Exact internal controllability of a one-dimensional aeroelastic plate*, Appl. Math. Optim., 24 (1991), pp. 99–111.
- [8] V. KOMORNIK, *On the exact interior controllability of a Petrowski system*, preprint.
- [9] J. LAGNESE, *Boundary Stabilization of Thin Plates*, SIAM Studies in Appl. Math., Vol. 10, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1989.
- [10] I. LASIECKA, J. L. LIONS, AND R. TRIGGIANI, *Nonhomogeneous boundary value problems for second order hyperbolic operators*, J. Math. Pures Appl., 65 (1986) pp. 149–192.
- [11] I. LASIECKA AND R. TRIGGIANI, *Exact controllability of semi-linear abstract systems with application to waves and plates boundary control problems*, Appl. Math. Optim., 223 (1991), pp. 109–154.
- [12] J. L. LIONS, *Contrôlabilité exacte perturbations et stabilisation de systèmes distribués*, Tome 1, Masson, Paris, 1988.
- [13] J. L. LIONS AND E. MAGENES, *Nonhomogeneous boundary value problems and applications*, Vol. 1., Springer-Verlag, Berlin, New York, 1972.
- [14] E. ZUAZUA, *Contrôlabilité exacte en un temps arbitrairement petit de quelques modèles de plaques*, in *Contrôlabilité exacte perturbations et stabilisation de systèmes distribués*, Tome 1, Appendice 1, Masson, Paris, 1988.
- [15] C. ZUILY, *Uniqueness and Nonuniqueness in the Cauchy Problem*, Progress in Math., Vol. 33, Birkhäuser, Boston, Basel, Stuttgart, 1983.

SHARP SUFFICIENT CONDITIONS FOR THE OBSERVATION, CONTROL, AND STABILIZATION OF WAVES FROM THE BOUNDARY*

CLAUDE BARDOS†, GILLES LEBEAU‡, AND JEFFREY RAUCH§

Abstract. For the observation or control of solutions of second-order hyperbolic equation in $\mathbb{R}_t \times \Omega$, Ralston's construction of localized states [*Comm. Pure Appl. Math.*, 22 (1969), pp. 807–823] showed that it is necessary that the region of control meet every ray of geometric optics that has, at worst, transverse reflection at the boundary.

For problems in one space dimension, the method of characteristics shows that this condition is essentially sufficient. For problems on manifolds without boundary, the sufficiency was proved in [J. Rauch and M. Taylor, *Indiana Univ. Math. J.*, 24 (1974)]. The theorems regarding propagation of singularities [M. Taylor, *Comm. Pure Appl. Math.*, 28 (1975), pp. 457–478], [R. Melrose, *Acta Math.*, 147 (1981), pp. 149–236], [J. Sjöstrand, *Communications in Partial Differential Equations*, 1980, pp. 41–94] allows the extension of the latter argument to the problem of interior control [C. Bardos, G. Lebeau, and J. Rauch, *Rendiconti del Seminario Matematico, Università e Politecnico di Torino*, 1988, pp. 11–32].

In this paper, the sufficiency is proved for problems of control and observation from the boundary. For multidimensional problems, the region of control must meet each ray in a nondiffractive point, and a new microlocal lower bound on the trade of solutions at the boundary at gliding points is required.

This paper treats linear problems with variable coefficients and solutions of all Sobolev regularities. The regularity of the controls is precisely linked to the regularity of the solutions.

Key words. controllability, observability, stabilization, geometric optics, propagation of singularities, rays

AMS(MOS) subject classifications. 35B40, 35L20, 93D20, 93B05, 93B07, 93C20

1. Introduction. This paper is devoted to the analysis of the observability, control, and stabilization of the solutions of second-order hyperbolic partial differential equations. The results obtained for multidimensional problems are as precise as those previously known for one-dimensional problems for which the method of characteristics is an elementary and effective tool [Ru, § 3].

We treat linear equations. However, by the usual linearization process, controllability of linear problems yields local controllability for nonlinear problems [LM, § 6.1]. If we linearize about a nonconstant solution, the linear equation has variable coefficients. If we linearize about a nonequilibrium solution, the coefficients will depend on time. Treating such time-dependent problems requires many technical innovations.

For multidimensional problems, several methods have been employed and have yielded rather incomplete results. The most special method is the use of eigenfunction expansions in the rare cases when the eigenfunctions/eigenvalues are known with some precision. In those cases, we obtain useful information that provides examples/counterexamples to guide a general development [Ru, § 4].

A more flexible method is the classical energy method, which consists of multiplying the differential equation by artfully chosen differential operators applied to u (the multiplier) and integrating by parts with the hope that enough terms will have the correct sign that we can derive inequalities, which imply observability, control, or

* Received by the editors, August 21, 1989; accepted for publication (in revised form) May 31, 1991. This research was partially supported by L'Année nonlinéaire française, 1988.

† UFR de Mathématiques, Université de Paris 7, 75251 Paris, France.

‡ Département de Mathématiques, Université de Paris 11, 91405 Orsay, France.

§ Department of Mathematics, University of Michigan, Ann Arbor, Michigan 48109. The work of this author was partially supported by National Science Foundation grant DMS-86-01783.

stabilization. The construction of clever multipliers was intensively developed in the study of scattering theory. In the context of scattering theory, the work of Morawetz, Ludwig, Lax, and Phillips is notable. Finally, Morawetz, Ralston, and Strauss [MRS] were able to prove local energy decay for arbitrary nontrapping obstacles using a far-reaching refinement of the method. They realized that multiplying by first-order differential operators applied to u was not sufficiently flexible in dimensions greater than 2. A key idea of theirs was to find symbols $q(x, \xi)$ of operators with desirable positivity properties, then to approximate the symbols by polynomials in ξ (of high order) so the resulting multipliers $q_{\text{approx}}(x, D)u$ use local operators. Fortunately, or unfortunately, depending on your point of view, their result has an independent proof based on the analysis of the propagation of singularities of solutions. Those results use the nonpolynomial symbols, that is, pseudodifferential operators, in an essential way. The key steps were the analysis of diffracted rays by Melrose and Taylor, followed by the study of gliding rays and higher-order contact by Melrose and Anderson, then Melrose and Sjostrand. One disadvantage of using the methods of microlocal analysis is that the domains and coefficients are required to be quite smooth. On the other hand, variable coefficients cause no trouble for that method, something that is far from true of the methods based on differential multipliers.

On the control theory side, key contributors using multipliers are Lions, Lagnese, Chen, Ho, Lasieka, and Triggiani. Their results are described in detail in the recent book of Lions [Lio]. The methods have the virtues of being elementary and requiring little regularity of coefficients and boundaries. In the simplest cases, where the multipliers are related to the dilation and conformal invariance of the wave operator (multipliers introduced by Morawetz in the early 1960s), the desired inequalities come with explicit constants. When the multipliers are refined to fit the geometry of less special domains, lower-order terms appear in the estimates. These lower-order terms are eliminated by compactness arguments, and we lose the explicitness of the constants. The method adapts well to control and stabilization from a subset of the boundary (in contrast to control from the entire boundary), but does not work so well for control from open subsets of the interior. When it works, the method yields sufficient conditions that are rarely sharp neither in the minimal time needed for control nor for the size of the set on which control must take place (see the examples at end of the Introduction). For problems with complicated geometry and/or coefficients that vary in space and/or time, the results obtained are usually very weak or nonexistent.

Another method that has been applied to multidimensional problems rests on the Holmgren uniqueness theorem. This requires real analytic coefficients that have the property of unique continuation. It is rare in applications that we believe that knowledge of properties of the medium in one region determines the properties in others. One striking case where this is true is homogeneous media that corresponds to constant coefficients. Otherwise, the real analyticity assumption is unreasonable from the point of view of most applications. Nevertheless, this approach does give some insight concerning the time needed for controllability. The uniqueness theorem when applied to a suitable dual problem yields the fact that the achievable states for a control problem are dense. Thus the Holmgren theorem yields results of approximate controllability [Ru, Thm. 5.1]. Recently, Lions [Lio] has observed that if we choose a Hilbert space norm $\|\cdot\|_F$ whose vanishing leads to a uniqueness theorem, then there are exact controllability theorems in the dual F' . This Hilbert uniqueness method (HUM) converts the approximate controllability theorems into exact controllability. The central difficulty is then shifted to the description of F, F' . Furthermore, the choice of clever Hilbert structure so that F' is reasonable is an art form not unrelated to the problem

of choosing clever multipliers. In cases where we have unicity but where the geometric condition of our paper is violated, the dual space F' is not even a space of distributions (see [BLR1], [Ha2], and § 4, below).

Yet another approach to multidimensional problems, in this case with constant coefficients, is to use explicit solution formulas based on the Fourier and/or Radon transform. Littman [Lit1] obtains sharp results for control from the whole boundary in this way.

Finally, we come to the idea/method that motivates our analysis. Wave equations have solutions that are localized near curves $(t, x(t))$ in space-time. The curves are called rays, and typical rigorous results assert that for any $\varepsilon > 0$ and $T > 0$ there is a solution so that the fraction of the energy located at a distance greater than ε from the ray is smaller than $1 - \varepsilon$ for $0 \leq t \leq T$ [Ra1], [Ra2]. To be able to observe such solutions, it is clear that we must observe on at least one point of every ray. Similarly, it would be foolhardy to try to control a solution from a set on which it is negligibly small. Thus controls must be so placed so that there is a control on every ray (Fig. 1). Note that, in this and the other figures, the x -projection of the rays are drawn. This point of view is completely general. It is applicable to any situation where the governing equations have localized solutions. For example, the equations of linear vibrations of plates, although not hyperbolic, have so-called Gaussian beam solutions [A], [Ra2], [Le]. This discussion suggests the following general principle:

To control, observe, or stabilize solutions of hyperbolic partial differential equations, it is necessary that we observe or control at least one point of each ray of geometric optics.

In his outstanding survey article, Russell [Ru] observes that

In contrast with the relatively complete theory which we have seen to exist for hyperbolic equations involving only a single space variable . . . the control and observation for processes in multidimensional regions is quite primitive. This is due to the fact that the characteristic surfaces arising in such problems are nowhere near as constructively useful as in the one-dimensional case [Ru, p. 680].

We show that the rays of geometric optics serve as well in higher dimension as in one dimension. In dimension equal to one, the rays and the characteristic hypersurfaces coincide. In higher dimensions, characteristic surfaces are swept out (foliated) by rays, which are the more informative objects.

Unfortunately, in higher dimensions, the necessary analysis is not as elementary as in dimension one. However, the hardest parts have already been performed in the study of propagation of singularities discussed above.

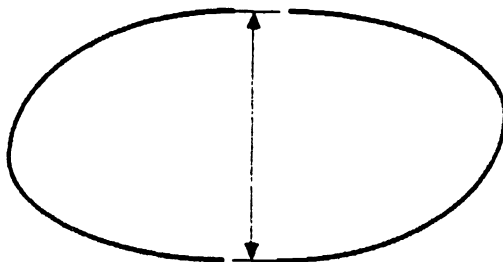


FIG. 1. With such a ray avoiding the region of action, exact controllability and stabilization cannot be achieved.

The use of rays in problems of observation, control, and stabilization was initiated by Rauch and Taylor [RT1], who studied stabilization and whose estimates are sufficient to prove controllability and observability in domains without boundary. Their point of departure was Hormander's theorem on the propagation of singularities in the absence of boundaries. They establish a three-step process, which we will follow. The three steps for studying observability are as follows.

Step 1. In the region where the observations are made, we have more information than elsewhere. Express that in the form of strong local/microlocal estimates.

Step 2. Use propagation of singularities theorems, which assert that effects propagate along rays together with the fact that every ray passes through the region of control to show that similar strong estimates are valid everywhere.

Step 3. The estimates from the use of propagation of singularities introduce "lower-order terms." Use a compactness argument to show that their effect is small or not present at all.

Relying only on Hormander's theorem on interior propagation together with Holmgren's theorem, Littman [Lit2] observed that we can obtain sharp results for control from the entire boundary. The condition for control then is that every ray hits $]0, T[\times \partial\Omega$. Littman shows that, if every ray crosses $]0, T[\times \partial\Omega$, then we can control from the entire boundary. In practice, the problem of controlling from a subset of the boundary is the most natural, and Littman's approach does not apply to that problem.

Combining the ideas from [RT1] with the results of Melrose, Taylor, Anderson, and Sjostrand, we obtain immediately and without any new technical tools precise results on the control and observation from open sets in the interior of a domain with boundary [BLR2].

For control from the boundary, Ralston [Ra2, § 5] used localized solutions to show the necessity of placing controls on all rays that are never tangent to the boundary. Recent advances allow the strengthening given in our Theorem 3.2.

To prove sufficiency for control from the boundary, new results concerning local effects of the boundary on waves are required to carry out Step 1. Microlocal analysis allows us to decompose any solution into a sum of terms, each microlocalized near a generalized bicharacteristic in the cotangent bundle. The projection of generalized bicharacteristics are called generalized rays.

Over the interior, generalized bicharacteristics are just the classical bicharacteristics of Hamilton–Jacobi theory. When they encounter the boundary transversely, they are reflected by the classical law of geometric optics. In this case, it is not difficult to show that the trace at the boundary is comparable in size to the corresponding wave. When a ray kisses the boundary at a diffractive point (Fig. 2), it is not hard to see that it is

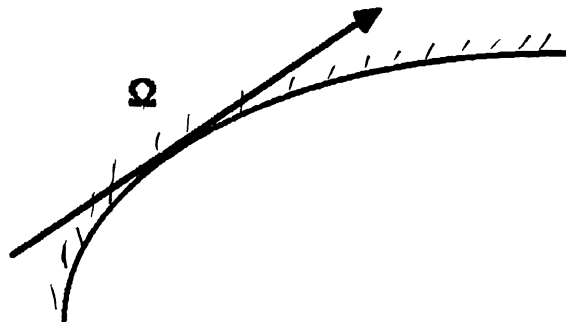


FIG. 2. A ray meeting the boundary at a diffractive point.

possible that it leaves a very small trace. The key inequality in this paper is the analysis of waves at near tangential incidence when the contact is not diffractive. A typical geometric situation is given by rays following regular polygonal paths on the inside of a disk in the limit when the number of sides tends to infinity. The limiting path hugs the circle and is called a gliding ray (Fig. 3). Our main estimate (given in “regularity form” in Theorem 2.2) implies that, for such gliding rays, the traces are comparable in size to the waves. This is reasonable since the rays, in the absence of the boundary, would proceed in a straight line; so the boundary must press on the wave to confine it inside the disk. This is in sharp contrast to diffractive rays whose direction is unaffected by the presence of the boundary.

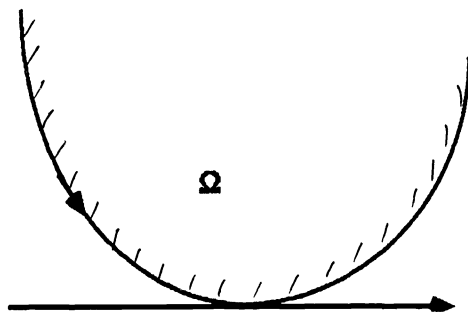


FIG. 3. *Gliding ray and nondiffractive point.*

The lower bound on traces, applied in a subset of the region of control or observation, is combined with the theorems on propagation of singularities to derive estimates throughout the domain.

We have written two expositions of our work, which were intended to introduce some of the ideas of microlocal analysis at the same time [BLR1], [BLR3]. The present exposition goes much further than these, in particular, treating time-variable coefficients and the full chain of regularities H^s , $s \in \mathbb{R}$. We have made an attempt to present results with natural hypotheses, which has forced us to generality and technicality in some cases. Furthermore, some incomplete arguments in [BLR1] are completed. Here we will use the microlocal analysis of boundary value problems as reported, for example, in [T2, Chap. IX] and [Ho, § 24]. The microlocal analysis presented in § 2 is greatly influenced by the methods of Melrose and Sjostrand [MS1]. The study of observability (§ 3) and stabilization (§ 5) follow the lines established in [RT1] for problems without boundary. Results on controllability in § 4 follow by a duality argument [Ru, § 2]. We have taken pains to treat carefully solutions of high regularity and the associated compatibility conditions at $\{t=0\} \times \partial\Omega$. The dual space to functions satisfying such conditions is not simple to describe but is the natural setting for the duality. This leads naturally to the study of mixed problems with distribution data where we follow the ideas of [RM]. Using these ideas we obtain controllability of solutions in a full scale of regularities with natural hypotheses.

Before passing to a more detailed description, we summarize some of the important features of our results.

Advantages.

1. The methods used work for operators with variable coefficients and arbitrary domains. The hypothesis depends on the relation between the rays of the operator and the subset on which the control is to take place.

2. The conditions obtained are very close to being necessary in the sense that slightly weaker conditions are necessary. For complicated operators and geometry, they may not be easy to verify, but there is no avoiding this because of the necessity.

3. The method works equally well for interior control, stabilization and control [BLR2], and combinations with boundary control.

4. The methods are flexible, having been used to treat related problems. Nalin [Na] has considered Maxwell's equations with natural boundary conditions; Lebeau has treated plate problems [Le1] and problems where the boundary has corners [Le2]. The methods have also been applied to pseudodifferential boundary conditions proposed as transparent boundary conditions in numerical analysis and as effective feedbacks for stabilization of the Dirichlet problem [BHLRZ].

Disadvantages.

1. The use of pseudodifferential operators requires the coefficients of the operator and the domain to be rather smooth. On the other hand, only a finite number of derivatives is needed. We have made no attempt to assess the number of derivatives. To do so in a serious way, we should take advantage of recent progress on pseudodifferential operators with irregular coefficients. To treat problems in H^s , the number of derivatives will grow linearly with s . We use a special case of the Malgrange preparation theorem, where an assessment of the required regularity would be less difficult. We believe that the number of derivatives needed is the sum of $|s|$ and a number between 2 and 10.

2. For problems with time-dependent coefficients, our general result shows that the achievable state is at most of finite codimension being the annihilator of the (finite-dimensional) family of *invisible* solutions. For such time-dependent problems, we must then examine as a separate question whether these states exist.

3. Another disadvantage is in the nature of the analysis. The use of geometric optics controls high-frequency effects. A finite-dimensional family of smooth solutions is outside the scope of such ideas. It is somewhat surprising that for time-independent equations, a short compactness argument together with unique continuation properties of the generator suffice to show that the finite space of invisible solutions is empty.

To give the flavor of our results, we describe a very special case. Suppose that $\Omega \subset \mathbb{R}^n$ is a bounded, open, connected subset such that $\bar{\Omega}$ is C^∞ embedded manifold with boundary. In $\mathbb{R}_t \times \Omega$ waves propagate governed by the real Klein-Gordon equation $(\square + 1)u = 0$, where \square is the speed one d'Alembertian, $\square \equiv \partial_t^2 - \Delta$. We attempt to observe, control, or stabilize from a relatively open subset $\omega \subset \partial\Omega$. At points of the boundary where there is no intervention, the waves satisfy the Neumann boundary condition.

The problem of observability takes the following form. The wave moves freely everywhere, that is, $\partial_\nu u = 0$ on all of $\partial\Omega$. The observer measures $u|_{]0, T[\times \omega}$, and the goal is to determine u throughout Ω in such a way that the recovery map $u|_{]0, T[\times \omega} \mapsto u$ is continuous. A natural norm for u is the energy norm

$$\int_{\Omega} u_t^2 + |\nabla_x u|^2 + u^2 dx / 2 \equiv e.$$

The rays for $\square + 1$ in \mathbb{R}^{1+n} are straight lines moving at speed one. If an \mathbb{R}^{1+n} ray encounters the boundary $\mathbb{R} \times \partial\Omega$ at a point t, x the associated generalized ray in $\mathbb{R} \times \bar{\Omega}$ is a gliding ray if the contact is exactly of order two, and the freely moving ray lies outside Ω locally except at t, x .

Modulo some technical hypotheses, we can construct solutions of equation $(\square + 1)u = 0$ satisfying the Neuman condition and localized close to the projection of any

generalized ray on $\mathbb{R} \times \Omega$. Thus, if there is a generalized ray whose projection on $\mathbb{R} \times \bar{\Omega}$ avoids $]0, T[\times \omega$, then there is no hope for continuous recovery from observations on $]0, T[\times \omega$. On the other hand, if a generalized ray passes over $]0, T[\times \omega$ only in diffractive points, it is possible that it will leave a very weak trace and therefore not be observable. We are led to the idea that we should observe on a set $\Gamma =]0, T[\times \omega$, such that every generalized ray meets $]0, T[\times \omega$ in either a point of reflection or a gliding point. Assuming this, a special case of Theorem 3.8 is that there is a constant $c > 0$ such that for all finite energy solutions,

$$(1.1) \quad \|u\|_{H^1([0, T[\times \omega])}^2 \geq ce.$$

In particular, we have exact and continuous observability, assuming hypotheses that are very nearly necessary.

A duality argument shows that (1.1) is equivalent to a result of exact controllability. Precisely, for any desired finite energy state $(f, g) \in \dot{H}^1(\Omega)^1 \times H^0(\Omega) \equiv Y$, there is a control $g \in \dot{H}^1([0, T[\times \partial\Omega) \equiv X$, which vanishes outside Γ such that the solution to

$$(\square + 1)u = 0 \quad \text{in }]0, T[\times \Omega, \quad u = g \quad \text{on }]0, T[\times \Omega, \quad u(0) = u_t(0) = 0$$

satisfies $(u(T), u_t(T)) = (f, g)$. To see the connection with (1.1), let K be the operator that takes g to the state $(u(T), u_t(T))$ (this operator is called C in [Ru]). Then K is continuous from \dot{H}^1 to $\dot{H}^1 \times L^2$. To show that it is onto, we must show that the transposed operator is subbounded in the sense that $\|K'y'\| \leq c\|y'\|$ for all $y' \in Y'$. With a suitable identification of the dual spaces, this estimate is exactly (1.1).

The subboundedness shows that the map KK' is a coercive map of Y' to itself. Its invertibility implies that K is onto. Inverting the map KK' is the idea of HUM. If we start with an estimate rather than a uniqueness theorem, then the space Y of achievable states, and the space X of controls are dictated by the estimate.

To compute the control g , we solve the equation $KK'y' = h$, where h is the target state and $g \equiv K'y'$. The coercive equation $KK'y' = h$ can be treated numerically by standard techniques, e.g., conjugate gradients. The underlying quadratic form is $\|K'y'\|_X^2$. As we will see in §4, the computation of this norm for given $y' \in Y'$ requires the solution of an initial boundary value problem for $\square + 1$ on $[0, T] \times \Omega$. Thus each iteration of conjugate gradients requires the solution of such a problem. These computations are feasible but nontrivial. The same calculations are required in the numerical implementations of HUM, so some experience has already been gained [GLL].

To illustrate the controllability criterion, we apply it to some simple sets $\Omega \subset \mathbb{R}^2$ and compare with the standard results based on Morawetz's multipliers. The latter yield the following sufficient condition for controllability (see [Lio, Chap. 1, Thm. 5.1]). Take any point $x \in \mathbb{R}^n$. A point $y \in \partial\Omega$ is called an exit point if $\nu \cdot (y - x) > 0$, where ν is an outward normal to $\partial\Omega$ at y . Drawing straight line segments leaving x , the exit points on $\partial\Omega$ are the points where the rays cut the boundary from the inside toward the outside. If ω is taken to be the set of exit points with respect to a center x , and $T > 2 \max\{|x - y| : y \in \partial\Omega\}$, then we have controllability with controls on $]0, T[\times \omega$.

Consider first Ω equal to the unit disk. Taking x as the center of the disk gives an absolutely sharp result for control from the entire boundary; T must be taken larger than the diameter. More interesting is the choice of x outside and far from Ω . Then the exit set is slightly more than half of the boundary, and we see that we can control from a set whose length is as close to $\frac{1}{2}$ we like. For sets close to a semicircle, the T given is very large. Our theorem gives controllability with $T > 6$, since if ω contains a semicircle, then every ray hits ω in each interval of time of length 6. If ω is slightly larger than the upper half of a circle, then the longest delay occurs for rays that include

a segment parallel to and just below the equatorial diameter. Such rays avoid ω for nearly three diameter lengths. Such a ray does not hit the semicircle until almost six units of time have past.

In the disk, each diameter is the x projection of a ray, and the necessary condition of Ralston shows that at least one of the two endpoints of each diameter must lie in ω if we are to control from ω . Thus $\omega \cup (-\omega)$ contains the entire boundary so $|\omega| \geq \pi$.

The following question has circulated in the control community since these results were obtained: Are there sets other than slightly extended half-circumferences that suffice for control and have length as close to π as we like? We construct an example settling the question in the affirmative as follows. Take an open arc ω in the boundary that contains a half-circumference and let p denote the midpoint of ω (Fig. 4). We omit a small neighborhood of p . For a ray to miss $\omega \setminus p$, it must hit p . Consider the rays through p . The diameter misses $\omega \setminus p$ as does the equilateral triangle with vertex p . Let θ denote the union of two open arcs centered, respectively, at the antipodal of p and one of the other vertices of the equilateral triangle. It is not difficult to show that, if ε is sufficiently small and γ_ε is the closed arc centered at p with length less than ε , then $]0, T[\times (\omega \cup \theta) \setminus \gamma_\varepsilon$ satisfies the hypothesis of the theorem for T sufficiently large.

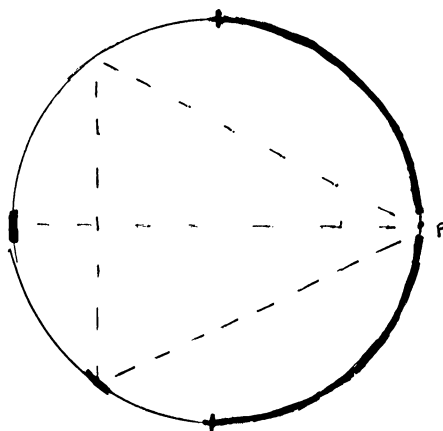


FIG. 4. A disconnected "minimal" region sufficient for control.

Another illuminating example is Ikawa's bowling ball, a disk with two disjoint interior disks removed (Fig. 5). Here every ray hits the boundary of the exterior disk with the exception of the ray that bounces back and forth between the two interior disks. If ω is the union of the boundary of the large disk and a small cap on one of the small disks containing one of the endpoints of the trapped ray, then $]0, T[\times \omega$ satisfies the hypothesis for T large.

A third example is the "dogbone" in Fig. 6. The first two diagrams show regions sufficient for control provable by the "exit criterion" above. The third set is also sufficient and is not easy to read from the standard results.

Our contention is not that we could not do any one of these three with a sufficiently clever differential multiplier. Quite the contrary, the methods of Morawetz, Ralston, and Strauss would surely suffice. However, to create a general result, we would be led inevitably to the same geometric considerations, and avoiding pseudodifferential techniques would only make the task more complicated.

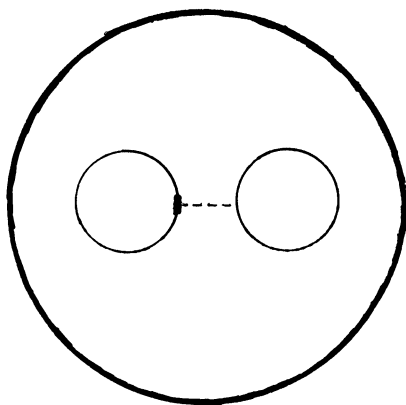


FIG. 5. Ikawa's bowling ball and a region sufficient for control.

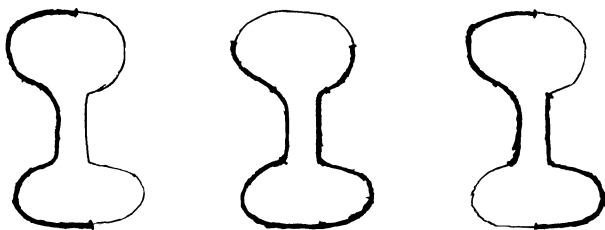


FIG. 6. Three regions sufficient for control in the dogbone region.

We turn next to problems of stabilization. Here the practical consideration is the addition of “passive devices” on the boundary to rapidly damp undesirable oscillations. A typical problem of stabilization occurs if the boundary condition is dissipative in the set Γ , and we ask if solutions must then decay exponentially to zero. Here the idea is to construct a system that is self-damping thanks to the dissipation in Γ . For example, consider the mixed problem

$$(\square + 1)u = 0 \quad \text{in }]0, t[\times \Omega, \quad (\partial_\nu + \alpha(x)\partial_t)u = 0 \quad \text{on }]0, T[\times \Omega,$$

where $\alpha \geq 0$ and $\omega = \{\alpha > 0\}$. The law of energy decay is

$$e(T) = e(0) - \int \int_{\Gamma} \alpha(x) u_t^2 d\sigma dt.$$

If α is not identically zero, then all solutions tend to zero [I]. On the other hand [Ra], if for every $T > 0$ there is a generalized ray that does not encounter $[0, T] \times \{\alpha > 0\}$, then there are solutions of finite energy whose energy decays to zero arbitrarily slowly. Exponential decay follows if we can show that there was a constant $c > 0$ so that

$$(1.2) \quad \int \int_{\Gamma} \alpha(x) u_t^2 d\sigma dt \geq ce(T).$$

In that case, the norm of the map from data at time zero to data at time T is at most $(1 + c)^{-1/2}$. If Γ satisfies exactly the same geometric condition as for the observability problem, we prove (1.2) in Theorem 5.5 where (1.2) is identical to (5.16).

2. A microlocal lower bound on traces. In addition to known results in linear wave propagation, in particular, the theorems of Taylor, Melrose, Anderson, and Sjostrand

on glancing and gliding rays, our study of control and observation requires a new result not unrelated to the above. This result asserts that, under suitable hypotheses, the trace at the boundary of a wave is comparable in size to the energy of the wave. Before giving the precise result, let us give a heuristic discussion that leads to the correct conclusions and hopefully renders the results intuitive.

Consider waves whose propagation in $M \equiv \mathbb{R}_t \times \Omega_x$ is governed by the d'Alembert wave equation together with homogeneous Dirichlet boundary conditions on $\mathbb{R}_t \times \partial\Omega$. The rays are then straight lines, corresponding to the fact that light rays travel along straight lines in empty space. The familiar laws of reflection apply at the boundary. At convex parts of the boundary, considering limits of rays approaching tangential incidence leads to the idea of gliding rays that travel along the boundary. Since a ray in free space would travel in a straight line, there must be a strong interaction between the boundary and the wave to bend a ray along a path gliding along the boundary. The boundary must push on the wave, and, by Newton's second law, the wave must push on the boundary. Thus we expect the trace of the wave at the boundary to be appreciable. Since the boundary is noncharacteristic and $u|_{\partial M} = 0$, the value of $\partial_\nu u$ at the boundary suffices to determine all the derivatives of u at the boundary. Thus we expect the trace of $\partial_\nu u$ at the boundary to be appreciable.

The same sort of argument applies to rays reflected at the boundary, since the direction of motion is abruptly changed at the boundary. The limiting case of tangential incidence is less clear and depends on the geometry of the region.

It is not difficult to make these impressions more precise in the case of a half-space $\Omega \equiv \{x_n > 0\}$. Write

$$x = (x_1, x_2, \dots, x_{n-1}, x_n) \equiv (x', x_n)$$

and

$$\xi = (\xi_1, \xi_2, \dots, \xi_{n-1}, \xi_n) \equiv (\xi', \xi_n).$$

Consider the reflected plane wave solutions

$$(\sin x_n \xi_n) \exp(i\xi' x' \pm i\tau t), \quad \tau = |\xi|.$$

The codirection τ, ξ' represents the direction of oscillation in the boundary. Note the decomposition of the codirections τ, ξ' into three regions, $\mathcal{H} \equiv \{\tau^2 > |\xi'|^2\}$, $\mathcal{G} \equiv \{\tau^2 = |\xi'|^2\}$, and $\mathcal{E} \equiv \{\tau^2 < |\xi'|^2\}$. For points in the hyperbolic region, \mathcal{H} , there are two plane wave solutions of the wave equation, $\exp(i(\xi_n x_n + \xi' x' + \tau t))$, with $\xi_n = \pm(\tau^2 - |\xi'|^2)^{1/2}$. In the glancing region \mathcal{G} there is one, and there are none for (τ, ξ') belonging to the elliptic region \mathcal{E} .

The angle of incidence θ between the direction of motion, ξ , and the normal to $\partial\Omega$, $(0, \dots, 0, 1)$, satisfies

$$\cos \theta = \xi_n / |\xi|.$$

Consider solutions u

$$u = \int A(\xi', \xi_n) (\sin x_n \xi_n) \exp(i\xi' x' + i|\xi|t) d\xi' d\xi_n.$$

The energy of u is proportional to

$$\int |\xi A(\xi', \xi_n)|^2 d\xi' d\xi_n.$$

The gradient of u on the boundary is equal to

$$\partial_n u|_{x_n=0} = \int \xi_n A(\xi', \xi_n) \exp(i\xi' x' + i|\xi|t) d\xi' d\xi_n.$$

Suppose that A is supported in $\xi_n > 0$ and make the change of variable $\xi', \xi_n \mapsto \tau, \xi'$ defined by $\tau = |\xi|$, $\xi_n = (\tau^2 - |\xi'|^2)^{1/2}$. Then

$$d\xi' d\tau = (\partial\tau/\partial\xi_n) d\xi' d\xi_n = (\xi_n/|\xi|) d\xi' d\xi_n$$

so

$$\partial_n u|_{x_n=0} = \int |\xi| A(\xi', (\tau^2 - |\xi'|^2)^{1/2}) \exp(ix'\xi' + i\tau t) d\xi' d\tau.$$

Plancherel's theorem in the x', t variables implies that the L^2 -norm of $\partial_n u$ on the boundary is proportional to

$$\int |\xi A(\xi', (\tau^2 - |\xi'|^2)^{1/2})|^2 d\xi' d\tau = \int |\xi A(\xi', \xi_n)|^2 (\xi_n/|\xi|) d\xi' d\xi_n.$$

If A is supported away from tangential incidence, the L^2 -norm of $\partial_n u$ and the energy are comparable. On the other hand, choosing A with support where $\xi_n/|\xi|$ is small, we can construct examples whose energy is arbitrarily large with respect to the L^2 -norm of $\partial_n u$. Even more, we can find solutions of infinite energy with $\partial_n u$ square integrable. The relation between the trace on the boundary and the energy from the frequency ξ is summarized by the heuristic principle

$$(2.1) \quad \frac{\text{Intensity of trace}}{\text{Intensity of wave}} \approx \cos \theta = \xi_n/|\xi|.$$

Turning to the general situation, suppose that $P(t, x, D_t, D_x)$ has smooth coefficients on \mathbb{R}^{n+1} and

$$P \equiv \partial_t^2 - \sum a_{ij}(t, x) \partial_i \partial_j + \text{lower order terms},$$

where a_{ij} is a symmetric real positive definite matrix on $\mathbb{R}_{t,x}^{n+1}$. Consider wave equations in $M \equiv \mathbb{R}_t \times \Omega$ where $\bar{\Omega}$ is a smooth embedded manifold with boundary in \mathbb{R}^n . Note that ∂M is noncharacteristic for P .

The symbol T is used to denote the tangent bundle, so, for example, the fiber T_y is the set of tangent vectors at y . The canonical projection from the tangent bundle to the base space is denoted by π . Similarly, T^* , T_y^* , and π are the cotangent bundle, its fiber at y , and the canonical projection. Recall that T_y^* is the dual of T_y .

A bicharacteristic is an integral curve in $T^*(\mathbb{R}^{n+1})$ of the Hamiltonian vector field H_p along which $p = 0$. Here

$$p(t, x, \tau, \xi) = \tau^2 - \sum a_{ij}(x) \xi_i \xi_j$$

is the principal symbol of P , and

$$H_p = (\partial H / \partial \tau) \partial / \partial t - (\partial H / \partial t) \partial / \partial \tau + \sum [(\partial H / \partial \xi_j) \partial / \partial x_j - (\partial H / \partial x_j) \partial / \partial \xi_j].$$

The projection of a bicharacteristic on t, x space is called a *ray*. We often concentrate on the x -projection, ignoring the time parameter.

For $y \in M$, $C^\infty(y)$ denotes the set of distributions u that are C^∞ on a neighborhood of y , that is, for which there is a $v \in C^\infty(\mathbb{R}^{1+n})$ such that $u = v$ on a neighborhood of y . Similarly, $H^s(y)$ denotes the set of distributions that are in the Sobolev space H^s on a neighborhood of y , that is, for which there is a $v \in H^s(\mathbb{R}^{1+n})$ such that $u = v$ on a neighborhood of y .

For $(y, \eta) \in T^*(M)$, $H^s(y, \eta)$ denotes the distribution that lie in H^s microlocally at y, η , that is, for which there is a $v \in H^s(\mathbb{R}^{1+n})$ such that $y, \eta \notin \text{WF}(u - v)$, where, as usual, WF denotes the wavefront set.

For $y \in \partial M$, $C^\infty(y)$ denotes the set of distributions u defined on a neighborhood of y in M that has the property that there is an $r > 0$, so on $M \cap B_r(y)$ u is the restriction of a $C^\infty(\mathbb{R}^{1+n})$ function. $H^s(y)$ for y in the boundary is defined by replacing C^∞ by H^s in the above definition.

We follow the definition of Chazarain (see [C], [MS1]) for microlocal regularity at the boundary. To avoid confusion with interior microlocal regularity, we will employ the notation H^s_{Ch} . This notion of microlocal regularity is the one used by Melrose and Sjostrand. In addition, for solutions of $Pu \in C^\infty$, it agrees with the intrinsic notion of Melrose [Me] (see [Ho, Cor. 18.3.33]), which does not depend on P .

DEFINITION. Suppose that $q \in T^*(\partial M)$ and u is a distribution defined on $B_r(\pi(q)) \cap M$ and satisfying $Pu \in H^{s-1}(B_r(\pi(q)) \cap M)$ for r sufficiently small. Then $u \in H^s_{Ch}(q)$ if and only if there is (1) a local change of variables $y = (y_0, \dots, y_n) = (y', y_n) = y(t, x)$ so that in the new variables $M = \{y_n > 0\}$, and (2) a tangential pseudodifferential operator $A(y, D_0, \dots, D_{n-1})$ of order zero such that A is elliptic at q and $Au \in H^s(B_r(\pi(q)) \cap \{y_n > 0\})$.

Note that near $\pi(q)$ we may subtract a solution $v \in H^s$ to $Pv = Pu$ reducing the above to the case where $Pu \in C^\infty$. In that case, the invariance of the definition is discussed in the above references.

The microlocal analysis at points $q \in T^*(\partial M)$ depends on whether q is hyperbolic, glancing, or elliptic according to the next definition. The natural inclusion $i: \partial M \hookrightarrow \mathbb{R}^{1+n}$ induces a map $i^*: T^*(\mathbb{R}^{1+n}) \rightarrow T^*(\partial M)$ such that i^* maps the conormal variety of ∂M to the zero section of $T^*(\partial M)$ and for $q \in T^*(\partial M)$, $(i^*)^{-1}(q)$ is a straight line in $T^*(\mathbb{R}^{1+n})$, parallel to the conormal to ∂M at $\pi(q)$.

DEFINITION. A point $q \in T^*(\partial M) \setminus 0$ is hyperbolic, glancing, or elliptic for P when $(i^*)^{-1}(q) \cap \text{char}(P)$ contains two, one, or no points. The set of hyperbolic, glancing, and elliptic points are denoted $\mathcal{H}, \mathcal{G}, \mathcal{E}$.

The next result makes precise the intuition about waves with nontangential incidence. The nontangential hypothesis takes the form $q \notin \mathcal{G}$. The result is stated as a regularity theorem. A quantitative version is valid (see Remark 4 following the proof of Theorem 2.2.).

THEOREM 2.1. Suppose that $q \in T^*(\partial M) \setminus \text{char}(P)$ and that u is a distribution defined on $B_r(\pi(q)) \cap M$ for r small positive and satisfying

$$Pu \in C^\infty(\pi(q)), \quad u|_{\partial M} \in H^s(q), \quad \partial_\nu u|_{\partial M} \in H^{s-1}(q).$$

Then $u \in H^s_{Ch}(q)$.

The reader is reminded that the hypotheses $u|_{\partial M} \in H^s(q)$ and $\partial_\nu u|_{\partial M} \in H^{s-1}(q)$ assert microlocal regularity.

Proof. The proof is straightforward starting from Taylor's decoupling (see [T2, Chap. IX]). Note that H^s regularity of u corresponds to H^{s-1} regularity of Taylor's w . We need only observe, in the notation of that reference, that if

$$w_j|_{y_n=0} \in H^{s-1}$$

and either

$$\partial_t w_j = i\lambda(y, D')w_j + a(y, D')w_j + C_0^\infty$$

or

$$\partial_t w_j = E_\pm(y, D')w_j + aw_j + C_0^\infty,$$

then $w_j \in H^{s-1}$ ($0 \leq y_n \leq 1$). For our second-order wave equations, the first alternative above occurs for the hyperbolic region, and the second in the elliptic region. The union of these two regions is $T^*(\partial M) \setminus \text{char}(P)$. The details are left to the reader. \square

For $q \in \mathcal{G}$, let \tilde{q} be the unique point in $\text{char}(P) \cap (i^*)^{-1}(q)$. Then $s \mapsto (\exp sH_p)\tilde{q}$ is the associated bicharacteristic (do not confuse s with that of the last paragraph) and the ray $\pi((\exp sH_p)\tilde{q})$ is tangent to ∂M . The point q is said to be *diffractive*, denoted $q \in \mathcal{G}_d$, if the ray has order of contact exactly equal to two with ∂M and lies in $\text{int}(M)$ for nonvanishing but small s . In this case the Taylor–Melrose theorem asserts that singularities of solutions follow the bicharacteristic. If the order of contact is two and the bicharacteristic lies over the exterior for small nonzero s , the point is said to be in the *gliding set* \mathcal{G}_g , and the Anderson–Melrose theorem asserts that singularities hug the boundary following gliding rays. The propagation of singularities for $\mathcal{G} \setminus (\mathcal{G}_d \cup \mathcal{G}_g)$ is treated by Melrose and Sjostrand. They introduced a uniform notation so that in all cases singularities propagate along *generalized bicharacteristics*.

Over the interior, $]0, T[\times \Omega$, Hormander's propagation of singularities theorem asserts that the wavefront set, in $T^*(]0, T[\times \Omega)$, is contained in $\{p = 0\}$ and is invariant under the flow of the Hamiltonian vector field H_p . There are solutions with wavefront set equal to any integral curve along which $p = 0$ (the *bicharacteristics*). In the same vein, there are solutions of $Pu = 0$ that are concentrated as close as we like to the projection of such bicharacteristics on t, x space. Such curves in space–time are called *rays of geometric optics* or simply *rays*.

For boundary value problems there is a natural way to extend bicharacteristics when they encounter the boundary, that is, when they pass over $[0, T] \times \partial\Omega$. We recall some of the central ideas and fix notation following [Ho, § 24.2], which contains a detailed discussion.

The simplest case is when they arrive transversally. Then there is a natural reflected bicharacteristic and reflected ray. Continuing bicharacteristics in this way leads to broken bicharacteristics and *reflected rays*. Ralston [Ra1] proved that there are solutions concentrated arbitrarily near any reflected ray.

If a bicharacteristic arrives tangent to $\partial T^*(\mathbb{R}_t \times \Omega)$ but with order of contact exactly two then the integral curve lies over the interior of Ω on a punctured neighborhood. In this case the extension is clear, and the projection on t, x space is called a *diffracted ray*.

Consider next the case of Ω equal to the unit disk in \mathbb{R}^2 and reflected rays with angle of incidence tending to zero. Passing to the limit, we find rays that hug the boundary, so-called *gliding rays*.

Points where a bicharacteristic has order of contact greater than two with the boundary can be transition points between a ray in the interior and one gliding along the boundary. In fact, if the field H_p has at most finite-order contact with the boundary, then every bicharacteristic has a unique continuation, leading to generalized bicharacteristic flow and *generalized rays*. In the case of infinite-order contact, a bicharacteristic may be continued in many ways as a generalized bicharacteristic [T1]. The set of continuations is a closed conic subset of $T^*(\mathbb{R} \times \bar{\Omega}) \setminus \{0\}$.

If i is the natural inclusion of ∂M into \mathbb{R}^{1+n} , then i^* gives a map from $T^*(\mathbb{R}^{1+n})|_{\partial M} \rightarrow T^*(\partial M)$. For $x \in \partial M$, replacing the points $q \in T_x^*(\mathbb{R}^{1+n})$ by $i^*(q)$ yields a map $T^*(\bar{M}) \mapsto T^*(M) \cup T^*(\partial M) \equiv T_b^*(M)$. The right-hand side is called the *compressed tangent bundle* and is given the quotient topology.

If γ is a generalized bicharacteristic and $\gamma(s)$ lies over a point of ∂M we associate to $\gamma(s)$ the point $i^*(\gamma(s)) \in T^*(\partial M)$. Making this replacement at all points over ∂M yields a curve $\tilde{\gamma}$ with values in $T_b^*(M) \setminus 0$. This curve is called the *compressed generalized bicharacteristic* associated to γ .

Our informal discussion suggested that gliding rays should leave an appreciable trace. This is further supported by the following heuristic analysis when Ω is the interior of the unit circle in \mathbb{R}^2 . Consider a ray reflecting at close to tangential incidence. The distance between successive reflections is $2(\cos \theta)$, where θ is the angle of incidence. Thus the number of reflections per unit length times the “impact per reflection” computed to be $\cos \theta$ in (2.1) remains constant. This suggests again the idea that the trace at the boundary of a gliding wave should be appreciable.

Diffraction rays have only a fleeting interaction with the boundary, and we might expect that there are examples for which the trace of the normal derivative at the boundary is small compared to the energy of the wave. This is also supported by the observation that such rays are the limit of rays reflected at the point of diffraction but with angle of incidence tending to zero. In contrast to the gliding case, there are no nearby reflections, so there is no accumulation of effects.

We will prove that the trace is appreciable for any ray that in the absence of boundary conditions would leave Ω . Such rays will be called *nondiffractive*. Special cases are the gliding and transversely reflected rays. As the boundary is responsible for confining the ray, it is reasonable to expect that the boundary must do appreciable work.

DEFINITION. A point $q \in T^*(\partial M) \setminus 0$ is called *nondiffractive* if (1) $q \in \mathcal{H}$, or (2) $q \in \mathcal{G}$ and the bicharacteristic $(\exp sH_p)\tilde{q}$ passes over the complement of M for arbitrarily small values of s , where \tilde{q} is the unique point in $\text{char}(P) \cap (i^*)^{-1}(q)$.

Examples. The following examples illustrate a variety of possibilities when $q \in \mathcal{G} \subset T^*(\partial M)$. In the examples, $n=2$ and a_{ij} is the identity matrix so that P_2 is the d'Alembert wave operator. Then bicharacteristics pass over straight lines in \mathbb{R}^{1+2} . The region Ω is the set $x_2 > f(x_1)$ with $f(0)=0=f'(0)$, so the x_1 -axis is tangent to $\partial\Omega$ at $(0,0)$. The glancing points in $T_{(0,0,0)}^*(\partial M)$ are multiples of $q_{\pm} \equiv (0,0,0;\pm 1,1)$ with $\tilde{q}_{\pm} = (0,0,0;\pm 1,1,0)$ and the bicharacteristics through these points pass over the same rays in \mathbb{R}^{1+2} traveling in opposite directions. Thus either both or neither are nondiffractive.

1. If $f''(0) > 0$, then Ω is convex near the origin and both $q_{\pm} \in \mathcal{G}_g$ are nondiffractive. Even more, they are gliding points.

2. If $f''(0) < 0$, then the complement of Ω is convex and both $q_{\pm} \in \mathcal{G}_d$ are diffractive, hence not nondiffractive.

3. If $f(s) = s^3$, q_{\pm} are nondiffractive. The bicharacteristic leaves \bar{M} as $\pm s$ increases and enters \bar{M} in the other sense.

4. If $f(s) = s^4$, both are nondiffractive.

5. If $f(s) = -s^4$, neither is nondiffractive. Note that neither is diffractive in the sense of Taylor and Melrose since the order of contact of the ray with $\partial\Omega$ is greater than two. Thus the nondiffractive points differ from the complement of the diffractive points.

6. If $f(x_1) = (\sin 1/x_1)(\exp -1/|x_1|)$, then q_{\pm} are nondiffractive, the rays passing in and out of M infinitely often near $(0,0,0)$.

The main result of this section is the following theorem.

THEOREM 2.2. *Suppose that $q \in \mathcal{G} \subset T^*(\partial M)$ is a nondiffractive point and that u is a distribution defined on $B_r(\pi(q)) \cap M$ for r small positive and satisfying*

$$Pu \in C^\infty(\pi(q)), \quad u|_{\partial M} \in H^s(q), \quad D_{t,x}u|_{\partial M} \in H^{s-1}(q).$$

Then $u \in H_{Ch}^s(q)$.

The case of C^∞ regularity, that is, $q \notin \text{WF}(u|_{\partial M}) \cup \text{WF}(\partial_\nu u|_{\partial M})$, implies that $q \notin \text{WF}_{Ch}(u)$ is Proposition 4.16 of [AM]. Their elegant argument does not suffice to prove Theorem 2.2.

Proof. The theorem is first proved for $s \leq 1$. The general case is derived from this by applying the $s \leq 1$ result suitable derivatives of u .

To treat the case where $s \leq 1$, we first reduce to the case where $u|_{\partial\Omega} = 0$. In our applications, we will have the stronger local (versus microlocal) regularity $u|_{\partial M} \in H^s(\pi(q))$ and $Du|_{\partial M} \in H^{s-1}(\pi(q))$. With these stronger assumptions, the reduction to the case where $u|_{\partial\Omega} = 0$ is by a simple localization (see [BLR1]).

Choose $\varphi_m \in C_0^\infty(\mathbb{R}^{n+1})$ with φ_m identically equal to one on a neighborhood of $\pi(q)$ and with supports shrinking to $\pi(q)$. Then, for m large, $\varphi_m Pu \in C^\infty(M)$.

For all m , $\text{WF}^s(\varphi_m u)/\mathbb{R}_+$ is a compact subset of the cosphere bundle $(T^*(M) \setminus 0)/\mathbb{R}_+$. For m large, $q \notin \text{WF}^s(\varphi_m u|_{\partial M})$.

Let $\tilde{\Gamma}$ denote the family of compressed generalized bicharacteristics in $T_b^*([0, T] \times \bar{\Omega})$ passing through q . Then $\tilde{\Gamma}/\mathbb{R}_+$ is a compact subset of $(T_b^*([0, T] \times \bar{\Omega}) \setminus 0)/\mathbb{R}_+$. Then $\tilde{\Gamma} \cap \text{WF}^s(\varphi_m u|_{\partial M})/\mathbb{R}_+$ is a decreasing family of compact subsets with empty intersection. Thus, for m large, $\tilde{\Gamma} \cap \text{WF}^s(\varphi_m u|_{\partial M}) = \emptyset$.

Let v be the solution of the initial boundary value problem

$$Pv = 0 \quad \text{in } M, \quad v|_{\partial M} = \varphi_m u|_{\partial M}, \quad \text{and} \quad v = 0 \quad \text{near } t = 0.$$

Write $\varphi_m u|_{\partial M} = g_1 + g_2$ with

$$g_1 \in H^s(\partial M) \cap \mathcal{E}'(\partial M) \quad \text{and} \quad \text{WF}(g_2) \cap \tilde{\Gamma} = \emptyset.$$

Corresponding to the decomposition of g , we have $v = v_1 + v_2$. Since the Dirichlet initial boundary value problem satisfies the strong estimates of Kreiss and Sakamoto, it follows that $v_1 \in H_{\text{loc}}^s(\partial M)$, and $Dv_1 \in H^{s-1}(\partial M)$. The Melrose-Sjostrand theorem [MS2] applied to v_2 implies that $q \notin \text{WF}_b(v_2)$. In particular, $v_2 \in H_{Ch}^s(q)$ and $Dv_2 \in H_{Ch}^{s-1}(q)$.

Subtracting v from u reduces to the case where $u|_{\partial M} = 0$ near $\pi(q)$.

Introduce local coordinates $y = (y_0, \dots, y_n) = (y', y_n)$ so that $M = \{y_n > 0\}$ and $\pi(q) = (0, 0)$. Following [MS1], the coordinates can be chosen so that

$$P = a(y)(D_n^2 + R(y, D')), \quad a \in C^\infty, \quad a(0) \neq 0.$$

Multiplying P on the left by a^{-1} , we may suppose that $a = 1$.

In these coordinates, the distribution u belongs to $C^\infty([0, \varepsilon[; \mathcal{D}'(|y'| < \rho))$, so it makes sense to extend u by zero in the complement of M . Denote this extension by \underline{u} . Then near $(0, 0)$,

$$(2.2) \quad P\underline{u} = \partial_n u(y', 0) \otimes \delta(y_n).$$

The main step in the proof is to show that $\underline{u} \in H^s(\tilde{q})$, where $\tilde{q} \equiv (i^*)^{-1}(q) \cap \text{char}(P)$, and i is the natural injection of ∂M into \mathbb{R}^{n+1} .

Since in ∂M , $Du \in H^{s-1}(\pi(q))$, we have $\partial_n u \otimes \delta \in H^{s-(3/2)-\varepsilon}(\pi(q))$ for all $\varepsilon > 0$. The $-\varepsilon$ can be omitted if $s < 1$. The microlocal elliptic regularity theorem (see [T2, Prop. VI.1.10]) implies that

$$(2.3) \quad \underline{u} \in H^{s+(1/2)-\varepsilon}(T_{\pi(q)}^*(\mathbb{R}^{n+1}) \setminus \text{char}(P)).$$

For $y \in \partial M$ and $(y, \zeta) \in (i^*)^{-1}(q) \in T_y^*(\mathbb{R}^{n+1})$ and not conormal to ∂M , $\partial_n u \in H^{s-1}(q)$ implies that $\partial_n u \otimes \delta \in H^{s-(3/2)}(y, \zeta)$. In particular, $\partial_n u \otimes \delta \in H^{s-(3/2)}(\tilde{q})$. Let Γ be the bicharacteristic through \tilde{q} . Then, near \tilde{q} , $Pu \in H^{s-(3/2)}(\Gamma)$. Since q is nondiffractive, Γ passes over the complement of M arbitrarily close to \tilde{q} and at such points u is smooth. Hormander's theorem (see [T2, Thm. VI.2.1]) shows that, near q , $u \in H^{s-(1/2)}(\Gamma)$. In particular, $u \in H^{s-(1/2)}(q)$. The crux of the argument is to improve this regularity to $u \in H^s(\tilde{q})$. Then we must show that u lies in $H_{Ch}^s(q)$.

The proof that $u \in H^s(\tilde{q})$ uses tangential pseudodifferential multipliers in the energy method is inspired by Melrose and Sjostrand [MS1]. The strategy is the following. Take the $L^2(y_n > 0)$ scalar product $(\cdot, \cdot)_M$ of Pu with Qu , $Q \equiv A_{2s-1}(y, D') + A_{2s-2}(y, D')D_n$, where the A_j are tangential pseudodifferential operators with real symbols homogeneous of degree j . The symbols of the A_j are supported in a small conic neighborhood of q .

Note that we then have $u \in C^\infty([0, \varepsilon[, \mathcal{D}'(|y'| < \rho))$, $Qu \in C^\infty([0, \varepsilon[, \mathcal{E}'(|y'| < \rho))$, and $Pu \in C^\infty([0, \varepsilon[: \mathcal{D}'(|y'| < \rho))$, so $(Qu, Pu)_M$ makes sense. The heart of our proof is using a carefully constructed Q so that integration by parts in this expression gives an estimate for $\|Cu\|_{L^2}$, where C is a pseudodifferential operator of order s elliptic at \tilde{q} . The integrations by parts are carried out assuming that u is smooth near $\pi(q)$. The justification in the present setting rests on the fact that u is the limit of a sequence u_m of regular solutions with uniform bounds on $\partial_n u_m$ in H^{s-1} microlocally at q . Lemma 4.7 below is a related result.

Integration by parts to move P from the right to the left, taking advantage of the fact that $u = 0$ on ∂M , yields [MS1, Cor. 2.6]

$$(2.4) \quad (Qu, Pu)_M - (QPu, u)_M = (u, ([P, Q] + (R^* - R)Q)u)_M.$$

As remarked above, the left-hand side is finite.

Now $[P, Q] + (R^* - R)Q$ is a sum of terms $E_{2s-j}(y, D')D_n^j$, $0 \leq j \leq 2$. Since $u(y', 0+) = 0$ in the support of the E_j , we see that

$$(2.5) \quad (u, ([P, Q] + (R^* - R)Q)u)_{\mathbb{R}^{n+1}} = (u, ([P, Q] + (R^* - R)Q)u)_M < \infty.$$

We will construct Q so that

$$\begin{aligned} [P, Q] + (R^* - R)Q &= C(y, D)^2 + \text{Op } S^{2s-2}(\mathbb{R}_y^{n+1} \times \mathbb{R}_{\xi, s}^n)P \\ &\quad + \text{Op } S^{2s-1}(\mathbb{R}^{n+1} \times \mathbb{R}^{n+1}), \end{aligned}$$

where $C(y, D)$ is of order s , has real symbol, and is elliptic at \tilde{q} . Since $Pu \in C^\infty$ and $u \in H^{s-1/2}(\tilde{q})$, it will follow that $u \in H^s(\tilde{q})$.

Note that for a pseudodifferential operator S , the principal symbol of $[P, S] + (R^* - R)S$ is equal to $H_\rho S + kS$, where $k(y, \zeta')$, homogeneous of degree 1 in ζ' , is the principal symbol of $R^* - R$ [MS1, (2.11)]. Note that if S is homogeneous of degree $2s-1$, then both the derivative $H_\rho S$ and the product $kS = k(y, \zeta')S(y, \zeta)$ are homogeneous of degree $2s$.

The strategy is to choose $C(y, \zeta)$ real homogeneous of degree s in ζ and supported in a small conic neighborhood of q in $\mathbb{R}^{n+1} \times \mathbb{R}^{n+1}$, and then to determine $S(y, \zeta)$ homogeneous of degree $2s-1$ in ζ by solving the transport equation

$$(2.6) \quad H_\rho S + kS = C^2$$

on a neighborhood of $y_n \geq 0$. This will guarantee that

$$[P, S(y, D)] + (R^* - R)S = C(y, D)^2 + \psi do(2s-1).$$

The term $H_p S$ is the derivative of S in the direction H_p . Thus (2.6) is a family of ordinary differential equations along the integral curves of H_p . The idea is to integrate (2.6) along those curves in the direction leaving M , with initial condition equal to zero before we reach the support of $C(y, \zeta)$. In the rare event of infinite-order contact of a ray with ∂M , there are technical problems, since bicharacteristics may leave and return infinitely often (Example 6, above, is an extreme case). In such cases the lines that follow are necessary. Otherwise simply follow Fig. 7. $Q(y, \zeta)$ will be constructed to be equal to $S(y, \zeta)$ modulo a multiple of $P(y, \zeta)$.

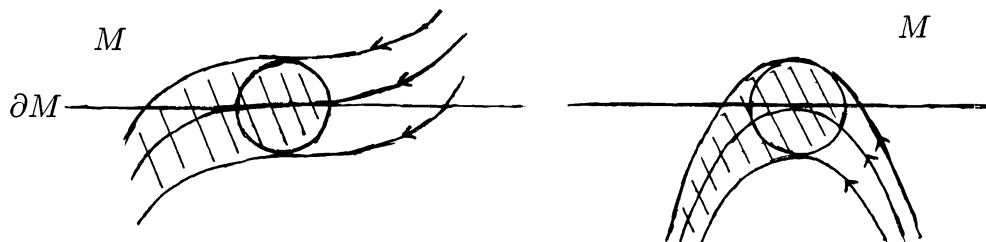


FIG. 7. The transport equation for $S(y, \zeta)$. The horizontal line is ∂M with M lying above. The curves, central disk, and shaded regions are the projections on t, y space of the bicharacteristics, $\text{supp}(C)$ and $\text{supp}(S)$, respectively. The order of contact is even in the figure on the right and odd in the figure on the left.

As q is nondiffractive, we may choose α arbitrarily close to zero so that $(\exp \alpha H_p) \tilde{q}$ lies over $y_n < 0$. We suppose that such α exist with $\alpha > 0$. The case where $\alpha < 0$ is treated similarly. Choose $\varepsilon_1 < 0 < \varepsilon_2$ so that

- (i) For $\alpha \in [\varepsilon_1, \varepsilon_2]$, $(\exp \alpha H_p) \tilde{q}$ lies over the coordinate patch we have chosen and does not pass through q for $\alpha \in [\varepsilon_1, \varepsilon_2] \setminus 0$;
- (ii) $(\exp \varepsilon_2 H_p) \tilde{q}$ lies over $y_n < 0$.

Next, choose a conic neighborhood of \mathcal{V} of \tilde{q} in $T^*(\mathbb{R}^{n+1})$ with the following properties:

- (i) For $\alpha \in [\varepsilon_1, \varepsilon_2]$, and $q \in \mathcal{V}$, $(\exp \alpha H_p) q$ lies over the coordinate patch we have chosen and does not intersect \mathcal{V} for $\alpha \in [\varepsilon_1, \varepsilon_2] \setminus [\varepsilon_1/2, \varepsilon_2/2]$;
- (ii) For $q \in \mathcal{V}$, $(\exp \varepsilon_2 H_p) q$ lies over $y_n < 0$.

Choose $C(y, \zeta)$ homogeneous of degree one in ζ , real, nonzero at \tilde{q} , and supported in \mathcal{V} .

Equation (2.6) is then solved as follows:

1. For bicharacteristics that do not pass through \mathcal{V} , S is set equal to zero.
2. For bicharacteristics through \mathcal{V} , the equation is integrated along the direction of the bicharacteristic $(\exp \alpha H_p) q$ with $S = 0$ for $\alpha \in [\varepsilon_1, \varepsilon_1/2]$. The integration is stopped at $\alpha = \varepsilon_2$. This is to avoid problems of reentry.

The symbol S is homogeneous of degree $2s - 1$, and the support of S over \bar{M} is a small conic neighborhood of the support of C . In particular, this support can be made as small as we like. Note that S is supported in a small conic neighborhood of q where $\zeta_n = 0 = R(y, \zeta)$.

The Malgrange preparation theorem [H, Thm. 7.5.6] allows us to divide S by $\zeta_n^2 + R(y, \zeta')$ to conclude that

$$(2.7) \quad S(y, \zeta) = A_{2s-1}(y, \zeta') + A_{2s-2}(y, \zeta') \zeta_n + A_{2s-3}(y, \zeta') (\zeta_n^2 + R(y, \zeta')),$$

where A_{2s-j} is homogeneous of degree $2s - j$ with respect to ζ' for $j = 1, 2$ and with respect to ζ for $j = 3$. Actually, the theorem applies only in a small neighborhood of \tilde{q} in $T^*(\mathbb{R}^{n+1})$. Applying the result only in $|\zeta| = 1$ and extending by homogeneity gives

(2.7) in a small conic neighborhood of \tilde{q} . Extend the resulting symbols A_{2s-j} to be smooth globally defined homogeneous symbols with support over our coordinate patch. Set

$$(2.8) \quad Q \equiv A_{2s-1}(y, D') + A_{2s-2}(y, D')D_n.$$

The constructions achieve the following identities microlocally at $\tilde{q} \in T^*(\mathbb{R}^{n+1})$:

$$\begin{aligned} S(y, D) &= Q + A_{2s-3}(y, D)P + \Psi_{2s-2}(y, D), \\ [P, Q] &= [P, S(y, D)] + [P, A_{2s-3}(y, D)P] + \Psi_{2s-1}(y, D), \\ [P, S(y, D)] + (R^* - R)S &= C(y, D)^*C(y, D) + \Psi_{2s-1}(y, D), \end{aligned}$$

where we are using the letter Ψ to denote pseudodifferential operators of the indicated order.

Next, return to the crucial estimate (2.5). The commutator of P and Q is equal to $[D_n^2 + R, A_{2s-1} + A_{2s-2}D_n]$. The $[D_n^2, A_{2s-2}D_n]$ term is equal to $\Psi_{2s-2}(y, D')D_n^2$. Replacing D_n^2 by $P - R$ yields an identity

$$[P, Q] + (R^* - R)Q = H_{2s}(y, D') + H_{2s-1}(y, D')D_n + \Psi_{2s-2}(y, D)P.$$

Define

$$G \equiv H_{2s}(y, D') + H_{2s-1}(y, D')D_n \equiv [P, Q] + (R^* - R)Q - \Psi_{2s-2}(y, D)P.$$

Then, by (2.5), $\infty > (u, Gu) = (u, \underline{Gu}) = (u, \underline{Gu})$.

Note that the essential support of the H_{2s-j} is close to q . Choose $\chi(y, D') = \chi(y, D')^*$ of order zero, supported near q , and identically equal to one on the essential support of the H_{2s-j} . Then the three quantities (u, Gu) , $(u, \chi G \chi u)$, and $(\chi u, G(\chi u))$ differ by bounded terms so they are all finite.

Both χu and $G(\chi u)$ have wavefront sets contained in a small conic neighborhood of $N^*(\partial M) \cup (i^*)^{-1}(q) \subset T^*(\mathbb{R}^{n+1})$. In addition, $(i^*)^{-1}(q) = \tilde{q} + N^*(\partial M)$. Next write $\chi u = v_1 + v_2 + v_3$ using an order zero pseudodifferential partition of unity $\{E_j(y, \xi): j = 1, 2, 3\}$ for WF (χu) so that $v_1 = E_1 \chi u$ has wavefront set near \tilde{q} , v_3 has wavefront set near $N^*(\partial M)$ and the wavefront set of v_2 is bounded away from $\text{char}(P) \cup N^*(\partial M)$. Then previously established regularity for u shows that

$$v_1 \in H^{s-(1/2)}, \quad v_2 \in H^{s+(1/2)}, \quad v_3 \in H^{s+(1/2)-\varepsilon}.$$

Write $(\chi u, G \chi u) = \sum (v_i, G v_j)$ and recall that $G = H_{2s} + H_{2s-1}D_n$. Thus, if neither i nor j is equal to 1, the summand is clearly bounded. If one of the indices is equal to one and the other is not, reason as follows. Choose \tilde{E}_j a partition of unity with slightly larger supports and with \tilde{E}_j equal to 1 on the essential support of E_j . Then

$$(v_i, G v_j) = (v_i, (\tilde{E}_i)^* G \tilde{E}_j v_j).$$

Now $(\tilde{E}_i)^* G \tilde{E}_j$ is a pseudodifferential operator of order $2s$ whose essential support is a subset of the region where both v_i and v_j belong to $H^{s+(1/2)}$. It follows that the summand is finite. The conclusion is that $(E_1 u, E_1 G u) < \infty$. In the computation that follows, we will drop the subscript 1 from E_1 .

$E(y, D)$ has essential support near \tilde{q} so that within the essential support of E the symbols S, P, C satisfy the relations established with the aid of the Malgrange preparation theorem. Thus

$$EG = E([P, Q] + (R^* - R)Q - \Psi_{2s-2}(y, D)P) = EC^*C + E\Psi_{2s-2}(y, D)P.$$

We know that $(Eu, EG u)$ is bounded. In addition,

$$(Eu, E\Psi_{2s-2}Pu) = (Eu, E\Psi_{2s-2}(\underline{Pu} + (\partial_\nu u) \otimes \delta(y_n))).$$

Since \underline{Pu} belongs to L^2_{loc} and $2s-2 \leq 0$, the $(E\underline{u}, E\Psi\underline{Pu})$ term is bounded. Next, $(\partial_\nu u) \otimes \delta(y_n) \in H^{s-(3/2)}$ on the essential support of E and $E\underline{u}$ belongs to $H^{s-(1/2)}$, which suffice to prove the boundedness of the second term. The conclusion is that $(E\underline{u}, EC^*Cu)$ is finite. This implies that $\underline{u} \in H^s(\tilde{q})$.

We next show that $u \in H^s_{Ch}(q)$. Since $u = \underline{u}$ in $y \geq 0$, it suffices to show that $A(y, D')\underline{u} \in H^1(\pi(q))$ for a tangential A , which is elliptic at q . With the aid of an order zero pseudodifferential partition of unity, write $\underline{u} = \sum \chi_j(y, D)\underline{u} + C^\infty$ near $\pi(q) = (0, 0)$. If the essential support of χ_j does not intersect $\text{char}(P)$, then (2.3) implies that $\chi_j(y, D)\underline{u} \in H^s(\pi(q))$. Therefore $A(y, D')\chi_j(y, D)\underline{u} \in H^s(\pi(q))$.

Since ∂M is noncharacteristic for M , we may suppose that for the other χ_j the essential support is disjoint from the conormal variety of ∂M . Then $A(y, D')\chi_j(y, D) \in \text{Op } S^0(\mathbb{R}^{n+1} \times \mathbb{R}^{n+1})$ that is a pseudodifferential operator in y . Its essential support belongs to $(i^*)^{-1}(\text{ess supp}(A))$. Choosing A with support in a sufficiently small neighborhood of q , the points of $\text{ess supp}(A\chi_j) \cap \text{char}(P)$ will lie in the elliptic set of C . Then $C\underline{u} \in H^{s-1}$ together with (2.3) yields $A\chi_j\underline{u} \in H^s(\pi(q))$.

Summing on j yields $A\underline{u} \in H^s(\pi(q))$, so \underline{u} and therefore u belong to $H^s_{Ch}(q)$. This completes the proof of the theorem for $s \leq 1$.

For general s , choose an integer m such that $s - m \leq 1$. For $\alpha \in \mathbb{N}^n$, $|\alpha| \leq m$, compute in the coordinates (y', y_n)

$$(2.9) \quad P((D')^\alpha u) = (D')^\alpha Pu + [R(y, D'), (D')^\alpha]u.$$

The commutator on the right is a differential operator of order $m+1$ with only D' derivatives, so it can be expressed in the form $\sum a_{\alpha,j}(y)D'_j D'^\alpha$. The first term on the right in (2.9) belongs to $H^{s-1-m}(\pi(q))$. Thus, for $U \equiv \{(D')^\alpha u\}_{|\alpha| \leq m}$, we have a large system $\mathcal{P}U \in H^{s-1-n}(\pi(q))$,

$$(2.10) \quad \mathcal{P} = \text{diag}(P, P, \dots, P) + PD'Op \text{ of order one.}$$

Furthermore, we have

$$(2.11) \quad U|_{\partial M} \in H^{s-m}(q) \quad \text{and} \quad DU|_{\partial M} \in H^{s-m-1}(q).$$

The proof we gave for scalar operators and $s \leq 1$ works with only routine modifications for systems of the form (2.10), (2.11). We conclude that $U \in H^{s-m}_{Ch}(q)$. This implies that $u \in H^s_{Ch}(q)$, and the proof is complete. \square

Remarks.

1. It is not hard to show that the nondiffractive hypothesis cannot be dispensed with. For example, at a diffractive point we can superpose solutions singular along rays that just miss the boundary to construct solutions satisfying all the other hypotheses of Theorem 2.2 but not the conclusion.

2. It would be natural to try to prove the theorem by reducing to the case where $u|_{\partial M} = 0$ and then writing the parametrix using Fourier-Airy operators. As the symbols lie in classes for which desirable L^2 -continuity properties may fail, we do not see how to carry out this argument.

3. More general operators P may be considered. In fact, we can achieve the generality of Melrose and Sjostrand. The cost here is that we can no longer subtract a solution to reduce to the case where $u|_{\partial M} = 0$, and we must analyze the equation $Pu = \partial_\nu u \otimes \delta_{\partial M} + u \otimes \delta'_{\partial M}$. This is, in fact, possible.

4. Tracing the steps of the demonstration, we find a quantitative version. That is, given that

$$E(y', D')\partial_\nu u(y', 0) \in L^2 \quad \text{and} \quad E(y', D')|D'|u(y', 0) \in L^2$$

with E is order $s-1$ and elliptic at q , we construct $A(y, D')$ of order s and elliptic at q and $B(y, D')$ of order zero and supported near q so that

$$\begin{aligned} \|A(y, D')u\|_{L^2} &\leq c(\|E(y', D')\partial_\nu u(y', 0)\|_{L^2} \\ &\quad + \|E(y', D')|D'|u(y', 0)\|_{L^2} + \|B(y, D')u\|_{L^2}), \end{aligned}$$

where the L^2 -norms are over $y_n \geq 0$.

3. Boundary observability of waves. In this section, we address the question of when observations at a subset of the boundary suffice to determine the values of u throughout the interior. We are interested in the case when this determination is a continuous function of the data. Such results are obtained by combining the results of the last section with the propagation of singularities theorems of Melrose-Sjostrand. Similar results for interior observations in problems without boundary were obtained in [RT]. Interior observations with boundary is discussed in [BLR2]. The strategy follows [RT].

In addition to the hypotheses of the last section, we suppose that

$$(3.1) \quad \begin{aligned} \bar{\Omega} \subset \mathbb{R}^n \text{ is an embedded compact connected} \\ \text{manifold with boundary.} \end{aligned}$$

In $]0, T[\times \Omega$ we suppose that

$$(3.2) \quad P(t, x, D_t, D_x)u = 0 \quad \text{in }]0, T[\times \Omega$$

with p as in § 2. At $\partial\Omega$, we impose the boundary condition

$$(3.3) \quad Bu = 0 \quad \text{on }]0, T[\times \partial\Omega.$$

On each component of $\partial\Omega$, B is a differential operator of degree zero or one with smooth coefficients, and ∂M is noncharacteristic for B . The order of B may vary from one component of $\partial\Omega$ to another. We suppose that

$$(3.4) \quad P, B \text{ is } L^2\text{-well-posed forward in time.}$$

The meaning of this is the following. For Cauchy data (u_0, u_1) and $t_1 \in [0, T]$, consider the initial value problem

$$\begin{aligned} Pu &= 0 \quad \text{on }]t_1, T[\times \Omega, \quad Bu = 0 \quad \text{on }]t_1, T[\times \Omega, \\ u(t_1, \cdot) &= u_0, \quad \text{and} \quad u_t(t_1, \cdot) = u_1. \end{aligned}$$

If $u_0, u_1 \in C^\infty(\bar{\Omega})$, then the initial value problem determines all the derivatives $D_{t,x}^\alpha u(t_1, x)$, $x \in \bar{\Omega}$. In particular, $\partial_t^j Bu(t_1, x)$ for $x \in \partial\Omega$ is determined by u_0, u_1 . A necessary condition for the existence of a solution smooth on $[t_1, T] \times \bar{\Omega}$ is that all these derivatives vanish. Let $CD_{P,B}^\infty(t_1)$ denote the set of Cauchy data for which such a smooth solution exists. The above remarks show that

$$CD_{P,B}^\infty(t_1) \subset \{(u_0, u_1) \in C^\infty(\bar{\Omega})^2: \partial_t^j Bu|_{\{t=t_1\} \times \partial\Omega} = 0, j=0, 1, \dots\},$$

where the derivatives $\partial_t^j Bu$ are determined by u_0, u_1 and the equation $Pu=0$. The Cauchy data on the right are said to satisfy the compatibility conditions to all orders (see [RM]).

DEFINITION. The operators P, B are L^2 -well-posed forward in time if, for every t_1 and all Cauchy data satisfying the compatibility conditions to all orders, there is a unique smooth solution of the mixed problem, and there is a constant c independent of $0 \leq t_1 \leq t_2 \leq T$ and u , such that the solution satisfies

$$\|u(t_2)\|_{H^1(\Omega)} + \|u_t(t_2)\|_{L^2(\Omega)} < c(\|u(t_1)\|_{H^1(\Omega)} + \|u_t(t_1)\|_{L^2(\Omega)}).$$

Cases of particular interest are Dirichlet's condition, $B_{\text{Dir}}u \equiv u$, and operators of Neumann type

$$(3.5) \quad B_{\text{Neum}} \equiv \partial_\nu + b(t, x),$$

where b is smooth on $\mathbb{R} \times \partial\Omega$ and ∂_ν is the normal derivative associated with P , that is,

$$(3.6) \quad \partial_\nu \equiv \sum a_{ij}(t, x) \nu_i(x) \partial_j$$

with ν denoting a smooth outward pointing conormal to $\partial\Omega$. The boundary operator is allowed to be of different type on different components of ∂M . The operator $B = \partial_\nu + a(t, x)\partial_t$ with $a \geq 0$ is well-posed forward in time, but if $d > 1$ and a is not identically zero then P, B is not L^2 -well-posed backward in time. Miyatake has given necessary and sufficient conditions for L^2 -wellposedness (see [Mi], [G]). Propagation of singularities for all boundary conditions satisfying (3.4) is treated by Melrose and Sjostrand [MS2].

We consider observations made on an open subset $\Gamma \subset]0, T[\times \partial\Omega$. Since the boundary is noncharacteristic for P , observation of the Cauchy data in Γ determines all the derivatives of u on Γ . Since we have one homogeneous boundary condition, it suffices to make one measurement. Thus, if $B = B_{\text{Dir}}$, we suppose that $\partial_\nu u|_\Gamma$ is observed, while if B is of order one, then $u|_\Gamma$ is observed.

Main question. Is the solution in $]0, T[\times \Omega$ determined by the observations on Γ , and if so, is the map from the values of u and $\nabla_{t,x}u$ on Γ to u in $]0, T[\times \Omega$ continuous?

Some information is provided by classical theorems. If the coefficients of P are real analytic, the Holmgren uniqueness theorem implies that observations on a neighborhood of a point (t, x) in Γ determine u on a neighborhood of (t, x) in $]0, T[\times \Omega$. When $n > 1$, $\Omega = \{x_n > 0\}$ is a half-space, and P has constant coefficients, Hadamard's test functions, $\exp(i\tau t + i\xi'x') \exp(\xi_n x_n)$ with τ, ξ' real and $\text{Re}(\xi_n) > 0$, show that this local determination is discontinuous.

The global Holmgren theorem of Fritz John yields global observability criteria (see [Ru, Thm. 5.2], [RS]). We give a representative result when $\Gamma =]0, T[\times \omega$ with $\omega \subset \partial\Omega$. For two points x and y in Ω , let $\text{dist}_\Omega(x, y)$ be the infimum of the lengths of smooth curves in Ω connecting x to y .

THEOREM 3.1. *Suppose that P has real analytic coefficients on $[0, T] \times \bar{\Omega}$ and that the local sound speeds of P are all greater than or equal to s . Let $L \equiv \sup \{\text{dist}_\Omega(x, y) : x \in \omega \text{ and } y \in \Omega\}$. If $T > 2L/s$, and $u \in \mathcal{D}'([0, T] \times \mathbb{R}^n)$ satisfies*

(i) $Pu = 0$ in $]0, T[\times \Omega$,

(ii) *The Cauchy data of u vanish on $]0, T[\times \omega$,*

then, u vanishes on $]0, T[\times \Omega$.

The fact that u is an extendable distribution and the boundary is noncharacteristic implies that the derivatives of u have well-defined limits in $\mathcal{D}'(\Gamma)$ the limits taken from inside $]0, T[\times \Omega$. In particular, the Cauchy data at Γ are well defined. Note that in Theorem 3.1, no boundary conditions at all are required outside of $]0, T[\times \Gamma$. This fact, the reliance on real analyticity, and Hadamard's construction should make us skeptical about the continuity of the reconstruction map. In fact, Theorem 3.1 proves uniqueness in many situations where reconstruction is not continuous (see Theorem 3.2 and the beginning of § 4).

Solutions of the boundary value problem (3.2), (3.4) describe waves with finite speed of propagation. Finite speed of propagation implies that we must observe on a sufficiently large set. For example, if the local sound speeds are always less than or equal to σ , then observation on $[0, T] \times \omega$ for $T < L/\sigma$ is insufficient to determine u . To see this, note that we may choose u with nonzero initial data supported in a small

neighborhood of a point $y \in \Omega$ with $\text{dist}(y, \omega) > T$. Such a solution will vanish on an $\mathbb{R} \times \bar{\Omega}$ neighborhood of $[0, T] \times \omega$.

By considering propagation of singularities instead of propagation of supports we find a necessary condition that is very nearly sufficient. The key objects are the rays of geometric optics.

In the generic case, the order of contact between a bicharacteristic and the boundary, $\partial(T^*(\mathbb{R} \times \Omega))$, never exceeds three. For real analytic operators and boundaries, the order is never infinite. The arguments of [Ho, § 24.3] and [Ta, Chap. IX] show that there are solutions of (3.2), (3.3) concentrated as close as we like to any generalized ray that has at most finite order of contact with the boundary. This immediately yields the following necessary condition.

THEOREM 3.2. *Suppose that P, B is L^2 -well-posed forward in time and that γ is a generalized bicharacteristic such that*

- (i) γ never passes over $[0, T] \times \bar{\Gamma}$,
 - (ii) at points of γ , H_p has at most finite order of contact with $\partial T^*(\mathbb{R} \times \Omega) \cap \{0 \leq t \leq T\}$.
- Then, for any $s \geq 0$, there is a sequence of solutions $u_k \in C^\infty([0, T] \times \bar{\Omega})$ to (3.2), (3.3) such that

- (i) $u_k|_{\Gamma}$ and $\partial_\nu u_k|_{\Gamma}$ converge to zero in $C_0^\infty[0, T] \times \Gamma$,
- (ii) $\|u_k\|_{H^s([0, T] \times \Omega)} = 1$.

This shows that to have continuous reconstruction, Γ must intersect the ray corresponding to every such γ . As indicated in the last section, waves can follow rays that just kiss ∂M without leaving an appreciable trace. It is not wise to try to observe such waves at the point of tangency. We should observe every ray at a nondiffractive point. At first, we suppose that infinite order contact does not occur,

$$(3.7) \quad H_p|_{\text{char}(P)} \text{ has at most finite order} \\ \text{contact with } \partial T^*(\mathbb{R} \times \Omega) \cap \{0 \leq t \leq T\}.$$

In this case, we say that *rays have finite-order contact with the boundary*.

THEOREM 3.3. *Suppose that P, B is L^2 -well-posed forward in time, that rays have finite-order contact with the boundary, and that $\Gamma \subset]0, T[\times \partial\Omega$ has the property that every compressed generalized bicharacteristic of P passes through a nondiffractive point in $T^*(\Gamma)$. Then there is an $\varepsilon > 0$ so that if $s \in \mathbb{R}$ and $u \in H^{s-1}(]0, T[\times \Omega)$ is a solution of (3.2), (3.3) with $D_{t,x}u|_{\Gamma} \in H^{s-1}(\Gamma)$, it follows that $u \in H^s(]T-\varepsilon, T[\times \Omega)$ and there are constants $c_1 > 0$ and c_2 such that*

$$(3.8) \quad \|\partial_\nu u\|_{H^{s-1}(\Gamma)} + \|u\|_{H^s(\Gamma)} \geq c_1 \|u\|_{H^s(]T-\varepsilon, T[\times \Omega)} - c_2 \|u\|_{H^{s-1}(]0, T[\times \Omega)}.$$

Proof. Let

$$X \equiv \{u \in H^{s-1}(]0, T[\times \Omega): (3.2), (3.3) \text{ hold and } D_{t,x}u|_{\Gamma} \in H^{s-1}(\Gamma)\}.$$

X is a Hilbert space with

$$\|u\|_X^2 \equiv \|u\|_{H^{s-1}(]0, T[\times \Omega)}^2 + \|Du\|_{H^{s-1}(\Gamma)}^2.$$

Combining Theorem 2.2 and the propagation of singularities, we will show that $X \subset H^s(\{T\} \times \bar{\Omega})$, that is, the elements of X are H^s at each point of the compact set $\{T\} \times \bar{\Omega}$. To do that, it suffices to show that for any $\rho \in T^*(M)$, with $t(\rho) = T$, $u \in H^s(\rho)$, and, for any $q \in T^*(\partial M)$ with $t(q) = T$, $u \in H_{ch}^s(q)$.

If $\rho \notin \text{char}(P)$ or if $q \notin \mathcal{H} \cup \mathcal{G}$, the desired regularity follows from microlocal elliptic regularity theorems (for example, [Ho, Thm. 20.1.14] treats the boundary case).

If $\rho \in \text{char}(P)$ (respectively, $q \in \mathcal{H} \cup \mathcal{G}$), then since (3.7) is satisfied, there is exactly one compressed generalized bicharacteristic passing through ρ (respectively, q). Denote by γ the corresponding generalized bicharacteristic. By hypothesis there is a nondiffractive point $q_1 \in T^*(\Gamma)$ belonging to the compressed bicharacteristic $\tilde{\gamma}$. By Theorem 2.2, the solution u belongs to $H_{ch}^s(q_1)$. Denote by γ_+ the part of γ with $t \geq t(q_1)$. We will show that u is microlocally H^s at each point of $\tilde{\gamma}_+$ by showing that u is the sum of two terms, $u = U_1 + U_2$, with $U_1 \in H^s(\]0, T[\times \Omega)$ and $\tilde{\gamma}_+ \cap \text{WF}_b(U_2) = \emptyset$.

Here WF_b denotes the wavefront set in the compressed cotangent bundle $T^*(M) \setminus 0 \cup T^*(\partial M) \setminus 0$ as defined in [MS1]. In that reference, $\text{WF}_b u$ is defined for extendable solutions of $Pu \in C^\infty(\bar{M})$. Unfortunately, our function U_2 does not satisfy $Pu \in C^\infty$, so the definition of $\text{WF}_b(U_2)$ requires care. We take advantage of the intrinsic theory of Melrose as described in [Ho, § 18.3]. The key result is that WF_b is well defined on a natural class of extendable distributions, denoted \mathcal{N} , which has the following properties:

(i) \mathcal{N} contains all extendable distributions that satisfy $Pu \in C^\infty(\bar{M})$ [Ho, Cor. 18.3.31];

(ii) \mathcal{N} is invariant under tangential pseudodifferential operators, and WF_b behaves in the natural fashion when such operators are applied [Ho, Thm. 18.3.32].

Choose $\chi(t) \in C^\infty(\mathbb{R})$ with $\chi = 0$ for $t < t(q_1) - \delta$ and $\chi = 1$ for $t \geq t(q_1)$. The small $\delta \in]0, t(q_1)[$ will be specified below. Let

$$f \equiv P(\chi u) \quad \text{and} \quad g \equiv B(\chi u).$$

Then f and g belong to \mathcal{N} , are supported in $t(q_1) - \delta < t \leq t(q_1)$ and for $t \geq t(q_1)$, u is equal to the solution U of

$$PU = f, \quad BU = g, \quad \text{and} \quad U = 0 \text{ for } t < t(q_1) - \delta.$$

The right-hand sides are written as the sum of two terms, $f = f_1 + f_2$ and $g = g_1 + g_2$. The first summands are equal to f and g microlocally at q_1 , and the f_j belong to \mathcal{N} . Toward that end, introduce local coordinates y such that $M = \{y_n > 0\}$, and the differential operator can be replaced by $D_n^2 + R(y, D')$. Abusing notation, we call the latter operator P .

Choose $a(y', \xi')$ homogeneous of degree zero in ξ' and supported in a small conic neighborhood of q_1 on which $u \in H_{ch}^s$, $u|_{\partial M} \in H^s$, and $\partial_n u|_{\partial M} \in H^{s-1}$. Choose $\varphi \in C^\infty(\mathbb{R})$ identically on a neighborhood of $\{0\}$ and supported in a very small neighborhood of that point. Let $A(y, D') \equiv \varphi(y_n)a(y', D')$. Choosing δ and the supports of a and φ small enough, we have $A(y', \xi')$ equal to one on a conic neighborhood of $\tilde{\gamma} \cap \{|t - t(q_1)| \leq \delta\}$, and, for any tangential pseudodifferential operator $\tilde{A}(y, F')$ with essential support contained in that neighborhood,

$$\tilde{A}u \in H^s(M), \quad \tilde{A}u|_{\partial M} \in H^s(\partial M), \quad \text{and} \quad \tilde{A}\partial_n u|_{\partial M} \in H^{s-1}(\partial M).$$

Let $f_1 \equiv Af$ and $g_1 \equiv Ag$. Then, with the supports chosen as above, $f, f_1 \in \mathcal{N}$ [Ho, Cor. 18.3.31 and Thm. 18.3.32] yields

$$\tilde{\gamma} \cap \text{WF}_b(f - f_1) = \emptyset \quad \text{and} \quad \tilde{\gamma} \cap \text{WF}_b(g - g_1) = \emptyset.$$

The fundamental result of Melrose-Sjostrand [MS2] extended to the \mathcal{N} category as in [Ho, Thm. 24.5.3] implies that $\tilde{\gamma} \cap \text{WF}_b(U_2) = \emptyset$.

On the other hand,

$$P\chi u = \chi Pu + [P, \chi]u = \text{DOp}_1 u + C^\infty,$$

so

$$f_1 = AP\chi u = \text{DOp}_1 Au + [A, \text{DOp}_1]u + C^\infty.$$

The first term belongs to H^{s-1} , and the commutator is a tangential pseudodifferential of degree zero with essential support contained in that of A , so the second term belongs to $H^s(\partial M)$. Thus $f_1 \in H^{s-1}(\partial M)$. Write $B = \alpha(y')\partial_n + \beta(y', D')$ with β a differential operator of degree one. Then

$$g_1 = AB\chi u = A\chi Bu + A[B, \chi]u = C^\infty + \tilde{A}(y', D')u,$$

where the essential support of \tilde{A} is contained in that of A , so the last term belongs to $H^{s-1}(\partial M)$ (H^s where $B = B_{\text{Dir}}$). Since P, B is L^2 -well-posed forward in time, it follows that $U_1 \in H^s(\cdot, T[\times \Omega)$.

This proves the desired decomposition of u , and we conclude that u is microlocally in $H^s(\rho)$ (respectively, $H^s_{\text{ch}}(q)$). Thus $u \in H^s(\{T\} \times \bar{\Omega})$, so there is an $\varepsilon > 0$ so that $u \in H^s(\cdot, T - \varepsilon, T[\times \Omega)$. In fact, the above proof produces an $\varepsilon > 0$ that is independent of $u \in X$.

The closed graph theorem applied to the inclusion $X \hookrightarrow H^s(\cdot, T - \varepsilon, T[\times \Omega)$ yields (3.8). \square

Turn next to the modifications that are needed when (3.7) is violated. The main difference is that we do not have unique generalized bicharacteristics passing through points of $T^*(\mathbb{R} \times \Omega)$. To show that u is microlocally regular at a point, we must consider a possibly infinite number of compressed generalized bicharacteristics terminating in that point. The simple compactness argument in the previous theorem must be buttressed somewhat.

Denote by \mathcal{B} , for bad, the set of glancing points in $T^*(\partial M)$ that are not nondiffractive. Thus $(T^*(\partial M) \setminus 0) \setminus \mathcal{B}$ is the set of nondiffractive points. The hypothesis of the last theorem is that every compressed bicharacteristic passes through a point of $T^*(\Gamma) \setminus \mathcal{B}$.

Denote by $\bar{\mathcal{B}}$ the closure of \mathcal{B} in $T^*(\partial M) \setminus 0$. The complement of $\bar{\mathcal{B}}$ in $T^*(\partial M) \setminus 0$ is the set of stably nondiffractive points, that is, the interior of the nondiffractive points. The hypothesis of the next theorem is that compressed bicharacteristics pass through points that are not only stably nondiffractive but stay away from the boundary of the stably nondiffractive points. Note that the hyperbolic points and the gliding points are stably nondiffractive.

In the cosphere bundle $(T^*(\partial M) \setminus 0)/\mathbb{R}_+$, $\bar{\mathcal{B}} \cap \{0 \leq t \leq T\}$ is compact. Identify the cosphere bundle with $\partial M \times S^n$, which is a metric space with the distance induced from $\partial M \times \mathbb{R}^{n+1}$.

THEOREM 3.4. *Suppose that P, B is L^2 -well-posed forward in time and $\Gamma \subset]0, T[\times \partial\Omega$ has the property that there is a conic neighborhood \mathcal{O} of $\bar{\mathcal{B}}$ such that every compressed generalized bicharacteristic of P passes through a point of $T^*(\Gamma) \setminus \mathcal{O}$. Then the conclusions of Theorem 3.3 are valid.*

Proof. Choose ε so small that the set of points in $(T^*(\Gamma) \setminus 0)/\mathbb{R}_+$ at distance less than or equal to 2ε from $\{0 \leq t \leq T\} \cap \bar{\mathcal{B}}/\mathbb{R}_+$ is contained in \mathcal{O} .

Let \mathcal{O}' be the set of points in $T^*(\Gamma) \setminus 0$ with image in $(T^*(\partial M) \setminus 0)/\mathbb{R}_+$, whose distance to $\{0 \leq t \leq T\} \cap \bar{\mathcal{B}}/\mathbb{R}_+$ is less than or equal to ε .

First, we show that there is an $m > 0$ such that every compressed generalized bicharacteristic passes through a point in $\{t \geq 1/m\} \cap T^*(\Gamma) \setminus \mathcal{O}'$. If not, there would be a sequence of generalized bicharacteristics γ_m , $m = 1, 2, \dots$, such that $\tilde{\gamma}_m$ never meets $\{t \geq 1/m\} \cap T^*(\Gamma) \setminus \mathcal{O}'$.

Let $T/2, x_m, \tau_m, \xi_m$ be the point over $t = T/2$ that belongs to γ_m . By homogeneity, we may suppose that $\tau_m^2 + |\xi_m|^2 = 1$. Passing to a subsequence, we may suppose that the γ_m converge to a limiting generalized bicharacteristic, γ . By hypothesis, there is a $t \in]0, T[$ such that $\tilde{\gamma}(t) \in T^*(\Gamma) \setminus \mathcal{O}$. By construction, $\gamma_m(t)$ converges to $\gamma(t)$.

Since $\gamma(t)$ is nondiffractive, the bicharacteristic $(\exp sH_p)\gamma(t)$ passes over the exterior of M for arbitrarily small s . Choose such an $s = \underline{s}$ so small that the closed interval I from t to \underline{s} is contained in $]0, T[$ and $(\exp sH_p)\gamma(t)$ passes over $T^*(\partial M)$ only at points of $T^*(\Gamma) \setminus \mathcal{O}$ for $s \in I$. It follows that for m large there is a $t_m \in I$ such that $\tilde{\gamma}_m(t_m) \in T^*(\Gamma) \setminus \mathcal{O}'$. In particular for m large γ_m passes over nondiffractive points of $T^*(\Gamma) \setminus \mathcal{O}'$ for $t > \min(t, s)$ violating the choice of γ_m .

As in the proof of Theorem 3.3, it suffices to show that for $s \leq 1$ (do not confuse with s of last paragraph)

$$\rho \in T^*(\mathbb{R} \times \Omega) \text{ with } t(\rho) = T \Rightarrow u \in H^s(\rho)$$

and

$$q \in T^*(\mathbb{R} \times \partial\Omega) \text{ with } t(q) = T \Rightarrow u \in H_{ch}^s(q).$$

For such points, let \mathcal{C} denote the set of compressed generalized bicharacteristics through ρ (respectively, q). For each $\tilde{\gamma} \in \mathcal{C}$, choose a point $q_\gamma \in \tilde{\gamma} \cap \{t \geq 1/m\} \cap T^*(\Gamma) \setminus \mathcal{O}'$. Denote by \mathcal{H} the closure of $\{\mathbb{R}_+ q_\gamma\}$ in $T^*(\partial M) \setminus 0$. Then $\mathcal{H} \subset \{t \geq 1/m\} \cap T^*(\Gamma) \setminus \mathcal{O}'$.

The decomposition performed in Theorem 3.3 with a compactness argument using the fact that \mathcal{H} is compact in the cosphere bundle expresses $u = u_1 + u_2$, with $u_1 \in H^s(]0, T[\times \Omega)$ and

$$Pu_2 = 0 \quad \text{in }]0, T[\times \Omega, \quad Bu_2 = 0 \quad \text{on }]0, T[\times \Omega, \quad \text{WF}_b(u_2) \cap \mathcal{H} = \emptyset.$$

In doing this, it is important that \mathcal{H} is contained in $\{t > 0\}$ since the evolution defined by P, B is only forward well posed.

If u were not in $H^s(\rho)$ (respectively, $H_{ch}^s(q)$), then ρ (respectively, q) would belong to $\text{WF}_b(u_2)$. Since the problem is forward well posed, the propagation of singularities theorem of Melrose and Sjostrand implies that there would be a bicharacteristic in \mathcal{C} that lies in $\text{WF}_b(u_2)$. In particular $\text{WF}_b(u_2) \cap \mathcal{H} \neq \emptyset$. This contradiction proves that $u \in H^s(\rho)$ (respectively, $H_{ch}^s(q)$). \square

The next result eliminates the hypothesis $u \in H^{s-1}$ from the previous two theorems.

COROLLARY 3.5. *If the hypotheses of Theorem 3.3 or Theorem 3.4 hold, then an extendable distribution, u , satisfying (3.2), (3.3) for $0 < t < T$ and $D_{t,x}u|_t \in H^{s-1}(\Gamma)$ must belong to $H^s(]T - \varepsilon, T[\times \Omega)$.*

Proof. Suppose that $\rho \in T^*([T - \varepsilon, T - \varepsilon/2] \times \Omega)$ (respectively, $q \in T^*([T - \varepsilon, T - \varepsilon/2] \times \partial\Omega)$). We will show that u is H^s at ρ (respectively q). Hypothesis (3.4) then yields the desired conclusion.

If ρ is not in the characteristic variety or q is not $\mathcal{H} \cup \mathcal{G}$, then the points do not belong to $\text{WF}_b u$, which is more than sufficient.

In the other cases, first suppose that (3.7) holds as in Theorem 3.3. Let $\tilde{\gamma}$ be the backward compressed generalized bicharacteristic passing through ρ or q . Then there is a nondiffractive $q_1 \in T^*(\Gamma)$ on $\tilde{\gamma}$. It is sufficient to show that u belongs to H^s along $\tilde{\gamma}_+$, the forward part of $\tilde{\gamma}$ from q_1 .

Since u is an extendable distribution, there is an $\underline{s} \in \mathbb{R}$ such that $u \in H^{\underline{s}}(\pi(q_1))$. The proof Theorem 3.3 then shows that $u \in H^{\min(\underline{s}+1, s)}(\gamma^+)$. By induction, we prove that $u \in H^{\min(\underline{s}+k, s)}(\gamma^+)$ for all k .

The modifications needed under the hypotheses of Theorem 3.4 resemble the proof of that theorem and are omitted. \square

DEFINITION. An extendable distribution u satisfying (3.2), (3.3) for $0 < t < T$ is called *invisible* if $u|_t = 0$ and $\partial_\nu u|_t = 0$. The set of all invisible solutions is denoted \mathcal{I} .

COROLLARY 3.6. *If the hypotheses of Theorem 3.3 or Theorem 3.4 hold, then the set of invisible solutions is a subspace of $C^\infty([T - \varepsilon, T] \times \bar{\Omega})$.*

Proof. Corollary 3.5 implies that u belongs to $H^s(\cdot]T - \varepsilon, T[\times \Omega)$ for all $s \in \mathbb{R}$. Thus $u \in C^\infty([0, T] \times \Omega)$. \square

To obtain more detailed information, we suppose in the remainder of the section that the boundary value problem P, B is reversable in the sense that

(3.9) P, B is L^2 -well-posed both forward and backward in time.

The operators of Dirichlet and Neumann type satisfy (3.9), while $\partial_\nu + a\partial_\mu$ does not when $d > 1$ and $a \geq 0$ is not identically zero. Necessary and sufficient conditions follow from Miyatake's characterization [Mi], [G].

A first consequence is that \mathcal{F} is contained in $C^\infty([0, T] \times \bar{\Omega})$, since solutions that are smooth near $t = T$ are smooth for $t < T$ because P, B defines a good backward evolution. In the same vein, solutions of (3.2), (3.3) that are H^s near $t = T$ are H^s on $]0, T[\times \Omega$. Thus the conclusions of Theorems 3.3 and 3.4 can be strengthened to $u \in H^s(\cdot]0, T[\times \Omega)$, and (3.8) can be strengthened to

$$(3.10) \quad \|\partial_\nu u\|_{H^{s-1}(\Gamma)} + \|u\|_{H^s(\Gamma)} \geq c_1 \|u\|_{H^s(\cdot]0, T[\times \Omega)} - c_2 \|u\|_{H^{s-1}(\cdot]0, T[\times \Omega)}.$$

COROLLARY 3.7. *If, in addition to the hypotheses of Theorem 3.3 or Theorem 3.4, P, B is L^2 -well-posed backward in time, then*

- (i) \mathcal{F} is a finite-dimensional subspace of $C^\infty([0, T] \times \bar{\Omega})$;
- (ii) *If u is an extendable distribution that satisfies (3.2), (3.3) for $0 < t < T$, $u|_\Gamma \in H^s(\Gamma)$, and $D_{\mu,s}u|_\Gamma \in H^{s-1}(\Gamma)$, then $u \in H^s(\cdot]0, T[\times \Omega)$ and the values of the traces determine the class of u in $H^s(\cdot]0, T[\times \Omega)/\mathcal{F}$;*
- (iii) *The map from the traces to H^s/\mathcal{F} is continuous.*

Proof. For invisible solutions, (3.10) shows that the $H^s(\cdot]0, T[\times \Omega)$ -norm of u is dominated by a constant times the $H^{s-1}(\cdot]0, T[\times \Omega)$ -norm. Since the imbedding of H^s in H^{s-1} is compact, Riesz's theorem implies that the set of invisible solutions is finite-dimensional.

Corollary 3.6 implies that the invisible solutions belong to $C^\infty(\cdot]T - \varepsilon, T[\times \bar{\Omega})$. Since P, B is well posed forward and backward in time it follows that $u \in C^\infty([0, T] \times \bar{\Omega})$.

Assertion (ii) is just the definition of \mathcal{F} .

To prove (iii), we must show that there is a constant $c > 0$ such that

$$(3.11) \quad \|\partial_\nu u\|_{H^{s-1}(\Gamma)} + \|u\|_{H^s(\Gamma)} \geq c \|u\|_{H^s(\cdot]0, T[\times \Omega)/\mathcal{F}}$$

for all extendable distribution solutions of (3.2), (3.3). The proof is indirect. If (3.11) were violated there would exist a sequence u_n for which the left-hand side tends to zero and $\|u_n\|_{H^s/\mathcal{F}} = 1$. Choose $i_n \in \mathcal{F}$ such that

$$(3.12) \quad \frac{1}{2} < \|u_n + i_n\|_{H^s} < \frac{3}{2}.$$

Estimate (3.10) applied to $u_n + i_n$ implies that

$$(3.13) \quad 0 < \liminf \|u_n + i_n\|_{H^{s-1}(\cdot]0, T[\times \Omega)}.$$

Since $u_n + i_n$ is bounded in H^s , Rellich's compactness theorem implies that we may select a subsequence, still denoted $u_n + i_n$, that converges strongly in $H^{s-1}(\cdot]0, T[\times \Omega)$ to a limit u . The limit is nonzero thanks to (3.13).

Since $u_n - i_n$ satisfies (3.2), (3.3), passage to the limit shows that the same is true of u . Then, $\partial_\nu(u_n + i_n)$ (respectively $(u_n + i_n)|_{\cdot]0, T[\times \partial\Omega})$) converge to $\partial_\nu u$ (respectively, $u|_{\cdot]0, T[\times \partial\Omega})$) in $H^{s-5/2}(\cdot]0, T[\times \partial\Omega)$ (respectively, $H^{s-3/2}(\cdot]0, T[\times \partial\Omega)$). Thus $u \in \mathcal{F}$.

Finally, applying (3.10) to the difference $u_n + i_n - u$, yields $u_n + i_n - u \rightarrow 0$ in $H^s(\cdot]0, T[\times \partial\Omega)$. This shows that u_n converges to zero in H^s/\mathcal{F} , contradicting (3.12). \square

Example. Let $d = 1$. $\Omega =]0, 1[$, $P = \square$, and $B = \partial_x + \partial_t$ for $x = 1$ and $B = B_{\text{Dir}}$ for $x = 0$. Waves moving to the right are not reflected at the right-hand boundary $x = 1$. Thus, if $\Gamma = [0, \infty[\times \{x = 0\}$, then initial data that launch no leftward-moving waves are invisible. Therefore the infinite-dimensional set of initial data satisfying $(\partial_t + \partial_x)u(0, x) = 0$ are invisible. This shows that a backward wellposedness hypothesis is needed.

In summary, the natural necessary condition for continuous recovery yields recovery modulo the finite-dimensional set of smooth invisible solutions. When Γ is of the form $]0, T[\times \omega$ corresponding to observations on a fixed subset independent of time, and the coefficients of P and B do not depend on time, the next theorem asserts that there are no invisible solutions, and we have continuous recovery.

THEOREM 3.8. *In addition to the hypotheses of Corollary 3.7, suppose that $\Gamma =]0, T[\times \omega$ and the coefficients of P and B do not depend on t . Then $\mathcal{J} = \{0\}$. In particular, observation of u and $D_{t,x}u$ on Γ determine uniquely solutions of (3.2), (3.3). The recovery map is continuous from $H^s(\Gamma) \times H^{s-1}(\Gamma)$ to $H^s(]0, T[\times \Omega)$.*

For some problems with real analytic coefficients, the conclusion $\mathcal{J} = \{0\}$ can be proved using Fritz John's global Holmgren theorem. This requires neither time-independent coefficients nor time-independent Γ .

Proof. We know that $\mathcal{J} \subset C^\infty$. The time independence hypotheses imply that if u belongs to \mathcal{J} then so does $\partial_t u$. Thus ∂_t is a linear map of the finite-dimensional space \mathcal{J} to itself.

If $\mathcal{J} \neq \{0\}$, this map must have an eigenvalue λ and nonzero eigenfunction u . Then $u_t = \lambda u$ in $]0, T[\times \Omega$, so $u = e^{\lambda t} v(x)$ for a $v \in C^\infty(\bar{\Omega})$. The differential equation satisfied by u implies that

$$\lambda^2 v - \sum a_{ij}(x) \partial_i \partial_j v + \sum a_i(x) \partial_i v + \lambda a_0(x) v = 0.$$

The fact that u is invisible implies that $v|_\omega = 0$ and $D_x v|_\omega = 0$. The unique continuation principle for second-order elliptic equations implies that v vanishes identically in Ω . It follows that $u \equiv 0$. This contradiction proves that $\mathcal{J} \equiv \{0\}$. \square

To compute the recovery map numerically, we can proceed as follows. For $s \in \mathbb{R}$, denote by X^s the set of H^s solutions to be observed

$$(3.14) \quad X^s \equiv \{u \in \mathcal{D}'(]0, T[\times \mathbb{R}^d): (3.2), (3.3) \text{ hold and } D_{t,x}u|_\Gamma \in H^{s-1}(\Gamma)\}.$$

Corollary 3.7 shows that X^s is closed in $H^s(]0, T[\times \Omega)$ and is a Hilbert space with norm

$$\|u\|_{X^s}^2 \equiv \|u\|_{H^s(\Gamma)}^2 + \|\nabla u\|_{H^{s-1}(\Gamma)}^2.$$

The observation map $O: X^s \rightarrow H^s(\Gamma)$ is defined by $u \mapsto u|_\Gamma$ for B of order one and $u \mapsto \partial_\nu u|_\Gamma$ for $B = B_{\text{Dir}}$. The recovery problem is to find u satisfying $Ou = g$ with g given in the range of O . When $\mathcal{J} = \{0\}$, (3.11) shows O^* is one-to-one on the range of O , so it suffices to solve $O^*Ou = O^*g$. This is equivalent to $(Ou, O\varphi)_X = (O\varphi, g)_X$ for all $\varphi \in X^s$, which can be taken as the starting point for a Galerkin method.

The problem of describing exactly the solutions with observed values of $D_{t,x}u|_\Gamma$ belonging to H^{s-1} is quite delicate. We end this section with two useful results in this regard. Define X^s as above. Then standard trace theorems in the Sobolev spaces $H^s(]0, T[\times \Omega)$ show that for $s > 3/2$

$$(3.15) \quad X^s \supset \{u \in H^{s+1/2}(]0, T[\times \Omega): (3.2), (3.3) \text{ holds}\}.$$

For smaller s , partial hypoellipticity at the boundary [Ho, Thm. B.2.9] yields the same result. The proof proceeds as follows. Use (3.9) to extend u to a solution of (3.2), (3.3) on $]a, b[\times \Omega$ with $a < 0 < T < b$. In local coordinates such that $M = \{x_1 > 0\}$ (this

variation of our usual notation is to agree with that of [Ho, App. B]. The regularity of u and Pu in the spaces $H_{(m,s)}(\mathbb{R}_d^+)$ of Hormander [Ho, App. B.2] is $u \in H_{(0,s)}^{\text{loc}}$ and $Pu \in H_{(\infty,0)}^{\text{loc}}$. Theorem B.2.9 yields $u \in H_{(\sigma,s-\sigma)}^{\text{loc}}$ for all $\sigma \in \mathbb{R}$. Take $\sigma = 1$ to see that $\partial_1 u \in H_{(0,s-1)}^{\text{loc}}$. For a boundary point \underline{x} in the coordinate neighborhood, choose a $\psi \in C_0^\infty(\mathbb{R}^{d+1})$ identically one near \underline{x} and supported in the neighborhood. Let $x = x_1, x'$ and $D = D_1, D'$. Then

$$\langle D' \rangle^{s-1/2}(\psi u) \in L^2([0, \infty[; H^{1/2}(\mathbb{R}_{x'}^d))$$

and

$$\langle D' \rangle^{s-(1/2)} D_1(\psi u) \in L^2([0, \infty[; H^{-1/2}(\mathbb{R}_{x'}^d)),$$

so $\psi u|_{\partial M} \in H^{s-(1/2)}(\mathbb{R}_{x'}^d)$. Thus $u|_{\partial M} \in H_{\text{loc}}^{s-(1/2)}$ and (3.15) follows.

In general, we cannot do much better than the sandwich $H^{s+(1/2)} \subset X^s \subset H^s$. However, if the boundary condition satisfies the uniform Lopatinski condition of Agmon–Kreiss–Sakamoto (see [CP], [Sak]), then H^s solutions satisfy automatically $D_{t,x} u|_{[0,T] \times \partial\Omega} \in H^{s-1}([0, T] \times \Omega)$. Of our simple examples, B_{Dir} and $\partial_\nu + a\partial_t$ with a nowhere zero satisfy the condition, while B_{Neum} does not. When the uniform Lopatinski condition is satisfied, it follows that the inclusion (3.15) is an equality. The above discussion is summarized in the next result.

COROLLARY 3.9. *If the hypotheses of Corollary 3.7 hold, then $X^s \subset H^s([0, T] \times \Omega)$ and (3.15) holds. If, in addition, B satisfies the condition of Agmon–Kreiss–Sakamoto, then, for all $s \in \mathbb{R}$,*

$$(3.16) \quad X^s = \{u \in H^s([0, T] \times \Omega) : (3.2), (3.3) \text{ holds}\}.$$

4. Boundary control of waves. This section is devoted to the exact controllability Theorem 4.9, which is dual to the results of the last section, particularly Corollary 3.7. We treat a scale of spaces covering a complete range of regularities. This requires a discussion of some technical questions related to compatibility conditions at the corner $\{t=0\} \times \partial\Omega$.

Suppose that P and B are as in (3.2), (3.3), and Γ is an open subset of $[0, T] \times \partial\Omega$. The basic problem is to steer a solution of $Pu = 0$ by way of controls exerted in the set Γ . Toward that end, consider a control $g \in \mathcal{D}'(\mathbb{R} \times \partial\Omega)$ with $\Gamma \supset \text{supp } g$ and solve the boundary value problem

$$(4.1) \quad Pu = 0 \quad \text{in } \mathbb{R} \times \Omega,$$

$$(4.2) \quad Bu = g \quad \text{on } \mathbb{R} \times \partial\Omega,$$

$$(4.3) \quad u = 0 \quad \text{for } t < 0.$$

The goal is to choose g so as to steer the solution so that the Cauchy data at time T , $(u(T), u_t(T))$ is a desired final state. This amounts to studying the map

$$(4.4) \quad K : g \mapsto (u(T), u_t(T)).$$

We are particularly interested in finding situations when this map is surjective or nearly so.

Increasing the space from which the controls g are taken or decreasing the size of the target space makes the task of exact controllability easier.

The existence of localized solutions, Theorem 3.2, showed that the rays of geometric optics play an essential role in the problem of observability. For controllability, it is the theorem of propagation of singularities that shows that the rays are crucial. If there

is a compressed bicharacteristic $\tilde{\gamma}$, which never passes over Γ and never makes infinite-order contact with the boundary, then, propagating regularity from the past, it follows that for any solution of (4.1)–(4.3), $\tilde{\gamma}$ is disjoint from $\text{WF}_b u$. Thus u is microlocally smooth along $\tilde{\gamma}$ and even microlocally real analytic if Ω , P , and B are real analytic [Sj]. In either case, the achievable states are infinitely smooth at $\tilde{\gamma} \cap \{t = T\}$. The conclusion is that if we want steerability to all states in $C^k([T, \infty[\times \bar{\Omega})$ for k sufficiently large, then we must control on a set large enough to encounter every ray of geometric optics.

This has an intriguing consequence for HUM. The space of achievable solutions is dual to the space F of Lions. Thus, if Γ misses $\tilde{\gamma}$, then F is not contained in a space of distributions of order $-k$ for any $k \in \mathbb{N}$. Even more striking, if P , B , and Ω are real analytic, then the space F contains elements that are not distributions. This has been observed in a special case by Haraux [Ha2].

For the basic problem to make sense, we must suppose that P , B generates a good time evolution forward in time. On the other hand, if P , B is not well posed backward in time, then signals may suffer an irreversible loss when they interact with the boundary. If we control before such an interaction, the effect of the control is likely to be lost. Thus, to achieve a state at time T , we are forced to control at all points of $\{t = T\} \times \partial\Omega$. This is dual to the fact that, for such systems, observability forces us to observe on the entire boundary at $t = 0$. If the boundary condition is irreversible only on a subset, we are forced to control on that subset at $t = T$. We will not pursue these ideas but assume that the boundary condition B is well posed both forward and backward in time.

We use the following criterion.

THEOREM 4.1 (Banach). *Suppose that X and Y are Banach spaces and that $K : X \rightarrow Y$ is a continuous linear map. Let $K' : Y' \rightarrow X'$ be the transpose. Suppose that $|\cdot|$ is a norm on Y' such that the identity map from Y' , $\|\cdot\|_{Y'}$ to Y' , $|\cdot|$ is compact. If there are constants $c_1 > 0$ and c_2 such that for all $y' \in Y'$*

$$(4.5) \quad \|K'y'\|_{X'} \geq c_1 \|y'\|_{Y'} - c_2 |y'|,$$

then the nullspace of K' is finite-dimensional, and the range of K is equal to the set of y annihilated by $\text{nullspace}(K')$.

Remarks.

1. In our applications, the spaces will be Hilbert spaces, and we use the word “transpose” to avoid confusion with the Hilbert space adjoint, which is very nearly the same object.

2. To apply this result, we must choose appropriate spaces X , Y . This choice is dictated by the inequalities of type (4.5), which we can prove for K' .

3. The best case is when $c_2 \leq 0$, in which case K is surjective.

To compute K' , consider K as a map from data in $C_0^\infty(\Gamma)$ to the Cauchy data at time T of smooth solutions, denoted $CD_{P,B}^\infty(T)$. All of our spaces of controls contain $C_0^\infty(\Gamma)$, and all of our target spaces are subsets of $CD_{P,B}^\infty(T)'$, so the desired transposes are all restrictions of the transpose of $K : C_0^\infty(\Gamma) \rightarrow CD_{P,B}^\infty(T)$.

To expose the essential features, we consider first the case of P equal to the D'Alembertian and $B = B_{\text{Dir}}$. In that case, the space $CD_{P,B}^\infty(T)$ is exactly $D(\Delta_D^\infty) \times D(\Delta_D^\infty)$, where

$$(4.6) \quad D(\Delta_D^\infty) \equiv \bigcap_{n \in \mathbb{N}} D(\Delta_D^n), \quad D(\Delta_D^n) = \{u \in H^{2n}(\Omega) : (\Delta^j u)|_{\partial\Omega} = 0 \text{ for } j \leq n-1\}.$$

If $\varphi_j(x)$ are real orthogonal normalized eigenfunctions of Δ_D arranged with eigenvalues nonincreasing, then $D(\Delta_D^\infty)$ consists of functions with rapidly decreasing Fourier

coefficients

$$u = \sum u_j \varphi_j, \quad \sup\{j^N |u_j| : j \in \mathbb{N}\} < \infty \quad \text{for all } N \in \mathbb{N}.$$

$D(\Delta_D^\infty)$ is a Fréchet space.

Elements of $C^\infty(\bar{\Omega})$ act as elements of the dual of $D(\Delta_D^\infty)$ by integration,

$$u \mapsto \langle u, \psi \rangle = \int u(x) \psi(x) dx = \sum u_j \psi_j,$$

where ψ_j are the Fourier coefficients of ψ . With this pairing, $D(\Delta_D^\infty)$ is weak star dense in $D(\Delta_D^\infty)'$ and the dual of $D(\Delta_D^\infty)$ is identified with the set of Fourier expansions with coefficients of polynomial growth.

Thus, to identify K' , it suffices to identify its action on $D(\Delta_D^\infty) \times D(\Delta_D^\infty)$. Toward that end, suppose that $(\psi_0, \psi_1) \in D(\Delta_D^\infty) \times D(\Delta_D^\infty)$ and consider

$$(4.7) \quad \langle Kg, (\psi_0, \psi_1) \rangle = \int u(T, x) \psi_0(x) + u_t(T, x) \psi_1(x) dx.$$

The key to the next computation is Green's identity

$$(4.8) \quad \begin{aligned} \int_{]0, T[\times \Omega} u \square \psi - \psi \square u dt dx &= \int_{\Omega} u \psi_t - u_t \psi dx \Big|_{t=0}^{t=T} \\ &+ \int_{]0, T[\times \partial \Omega} \psi \partial_\nu u - u \partial_\nu \psi dt d\sigma. \end{aligned}$$

Choose ψ as the solution to

$$(4.9) \quad \square \psi = 0, \quad \psi|_{\mathbb{R} \times \partial \Omega} = 0, \quad \psi(T) = -\psi_1, \quad \psi_t(T) = \psi_0.$$

This yields

$$(4.10) \quad \langle Kg, (\psi_0, \psi_1) \rangle = \int_{]0, T[\times \partial \Omega} g \partial_\nu \psi dt d\sigma.$$

This proves that $K'((\psi_0, \psi_1)) = \partial_\nu \psi|_\Gamma$, where ψ is the solution of (4.9). Here $\partial_\nu \psi \in C^\infty(\mathbb{R} \times \partial \Omega)$ acts as an element of the dual to $C_0^\infty(\Gamma)$ by integration with respect to $dt d\sigma$.

This result extends to general $(\psi_0, \psi_1) \in (D(\Delta_D^\infty) \times D(\Delta_D^\infty))'$ once the solution of (4.9) with such initial data is defined. In this simple case, we have the explicit formula

$$(4.11) \quad \begin{aligned} &(\cos((- \Delta_D)^{1/2}(t-T)))\psi_0 - (\sin((- \Delta_D)^{1/2}(t-T)))(- \Delta_D)^{-1/2}\psi_1 \\ &= \sum [(\psi_0)_j \cos((- \lambda_j)^{1/2}(t-T)) - (\psi_1)_j (\sin((- \lambda_j)^{1/2}(t-T))/(- \lambda_j)^{1/2})] \varphi_j. \end{aligned}$$

Since $\square \psi = 0$ in the sense of $\mathcal{D}'(\mathbb{R} \times \Omega)$ and $\mathbb{R} \times \partial \Omega$ is noncharacteristic, the trace $\partial_\nu \psi|_\Gamma$ is well defined, and identity (4.10) extends by continuity to all $g \in C_0^\infty(\Gamma)$ and $(\psi_0, \psi_1) \in (D(\Delta_D^\infty) \times D(\Delta_D^\infty))'$.

THEOREM 4.2. *Suppose that $P = \square$, $B = B_{\text{Dir}}$; then $K : C_0^\infty(\Gamma) \rightarrow D(\Delta_D^\infty) \times D(\Delta_D^\infty)$ defined in (4.1)–(4.4) has transpose given by $K'((\psi_0, \psi_1)) = \partial_\nu \psi|_\Gamma$, where ψ is the solution of (4.9) defined above.*

For general P, B , the identification is similar. Lost is the simple explicit formulas for $CD_{P,B}^\infty(T)$, $CD_{P,B}^\infty(T)'$. In addition to being smooth on $\bar{\Omega}$, the elements of CD^∞ satisfy an infinite set of compatibility conditions at $\partial \Omega$. When P and B are time dependent, these conditions may be time dependent. If the coefficients of P and B do not depend on time, then CD^∞ is time independent, and a description of intermediate complexity is possible (see [RM, Remark 1.2]).

As in Theorem 4.2, the description of K' involves the solution of an initial boundary value problem with initial data in the dual of $CD_{P,B}^\infty(T)$. The boundary condition on $\mathbb{R} \times \partial\Omega$ is the adjoint boundary condition, $B'u = 0$. The key properties of this condition are that P', B' is well posed with time running backward whenever P, B is forward well posed, and the evolution operator for P', B' gives the transpose of the evolution operator of P, B . If $B = B_{\text{Dir}}$, the adjoint boundary operator is also B_{Dir} . For B of order one with $\mathbb{R} \times \partial\Omega$ noncharacteristic, we may multiply by a nonzero smooth function on the boundary to replace B by

$$(4.12) \quad B = \partial_\nu + A_1(t, x, D_{t,x}) + A_0(t, x) \equiv \partial_\nu + A,$$

where A_1 is a vector field tangent to $\mathbb{R} \times \partial\Omega$, A_0 is a smooth function on the same set, and ∂_ν is given by (3.6). The A_i may be complex. A special role is played by time derivatives that may occur in A_1 , so we write

$$(4.13) \quad A_1 = a(t, x)\partial/\partial t + \text{terms in } \partial/\partial x_j.$$

For u and ψ smooth on $[0, T] \times \bar{\Omega}$, Green's identity reads

$$(4.14) \quad \int_{]0, T[\times \Omega} uP'\psi - \psi Pu \, dt \, dx = \int_{\Omega} u\psi_t - u_t\psi \, dx \Big|_{t=0}^{t=T} + \int_{]0, T[\times \partial\Omega} \psi \partial_\nu u - u \partial_\nu \psi \, dt \, d\sigma.$$

When u satisfies $(\partial_\nu + A)u = 0$, the boundary term becomes

$$- \int_{]0, T[\times \partial\Omega} \psi Au + u \partial_\nu \psi \, dt \, d\sigma.$$

Green's identity simplifies exactly when ψ satisfies $(\partial_\nu + A')\psi = 0$, where A' is the transpose of the operator A on $\mathbb{R} \times \partial\Omega$ with respect to the measure $dt \, d\sigma$. For B as in (4.12), the adjoint boundary operator is defined to be $B' \equiv \partial_\nu + A'$. If ψ satisfies the adjoint boundary condition $B'\psi = 0$ on $\mathbb{R} \times \partial\Omega$, we have

$$(4.15) \quad \int_{]0, T[\times \Omega} uP'\psi - \psi Pu \, dt \, dx = \int_{\Omega} u\psi_t - u_t\psi \, dx \Big|_{t=0}^{t=T} - \int_{\partial\Omega} au\psi \, d\sigma \Big|_{t=0}^{t=T}.$$

If $u \in \mathcal{D}'(\mathbb{R} \times \mathbb{R}^d)$ satisfies $Pu = 0$, then the derivatives of u have well-defined traces at the boundary, since it is noncharacteristic. There is no difficulty in defining solutions of $Pu = 0$, $Bu = g$ for $u \in \mathcal{D}'(\mathbb{R} \times \Omega)$ and $g \in \mathcal{D}'(\mathbb{R} \times \partial\Omega)$. If u and g vanish for $t < t_0$, such solutions are uniquely determined by g . The situation is not so simple for the initial value problem

$$(4.16) \quad Pu = 0 \quad \text{on }]0, T[\times \mathbb{R}, \quad Bu = 0 \quad \text{on }]0, T[\times \partial\Omega,$$

$$(4.17) \quad u(0, \cdot) = u_0(\cdot), \quad \text{and} \quad \partial_t u(0, \cdot) = u_1(0, \cdot) \quad \text{on } \Omega,$$

when the Cauchy data are distributions. The traces of u and u_t define distributions on Ω , but these distributions need not uniquely determine the solution. The difficulty comes from the presence of the corner $\{t = 0\} \times \partial\Omega$ and is discussed in [RM, § 5]. The action of the initial data must be given on a larger class of test functions than $\mathcal{D}(\Omega)$ to sense the values at $\partial\Omega$. In the example of \square , B_{Dir} the Fourier expansion method suggested the data be taken from $(CD^\infty)'$. The same result is correct in general.

We introduce some notation that aids the exploitation of (4.15). Define a bilinear form, $[\cdot, \cdot]$, on $C^\infty(\bar{\Omega}) \times C^\infty(\bar{\Omega})$ by

$$(4.18) \quad [(u_0, u_1), (\psi_0, \psi_1)] = \int_{\Omega} u_1 \psi_0 - u_0 \psi_1 \, dx + \int_{\partial\Omega} a u_0 \psi_0 \, d\sigma.$$

Denote by $U(t)$ the Cauchy data, $(u(t), u_t(t))$, and similarly $\Psi(t)$. If u is a smooth solution of $Pu = 0$, $Bu = 0$, and Ψ is the smooth solution of $P'\psi \in F \in \mathcal{D}'([0, T] \times \Omega)$, $B'\psi = 0$, $\psi = 0$ for $t \geq T$, then

$$(4.19) \quad \int_{\mathbb{R} \times \Omega} uF \, dt \, dx = [(u_0, u_1), (\psi_0, \psi_1)] = [U(0), \Psi(0)].$$

This suggests that $U(0)$ should be restricted to lie in a class of distributions so that the right-hand side of (4.19) makes sense. In this way, the initial data $U(0)$ are viewed as linear functionals on $CD_{P,B}^\infty(0)$. The next two propositions show how solutions with such data are constructed.

PROPOSITION 4.3. *The map $(u_0, u_1) \mapsto [(u_0, u_1), \cdot]$ injects $CD_{P,B}^\infty(t)$ onto a dense subset of $CD_{P',B'}^\infty(t)'$, where the latter space is given the weak star topology.*

Proof. If u_0, u_1 is sent to the zero linear functional, then for ψ_0, ψ_1 in $\mathcal{D}(\Omega)$ we have $\int u_0 \psi_1 - u_1 \psi_0 \, dx = 0$, so u_j vanish in Ω . Since $u_j \in C^\infty(\bar{\Omega})$, this suffices to show that $u_j = 0$.

To show that the image is dense, it suffices to show that there is no element (ψ_0, ψ_1) in $CD_{P',B'}^\infty(t)$ that is annihilated by the range. For such ψ we would have $[U, \Psi] = 0$ for all $U \in CD_{P,B}^\infty(0)$. Considering U supported in the interior of Ω yields $\psi_j = 0$ on Ω . Since $\psi_j \in C^\infty(\bar{\Omega})$, $\psi_j = 0$. \square

PROPOSITION 4.4. *For $T > 0$, let $S_{P,B}: CD_{P,B}^\infty(0) \rightarrow C^\infty([0, T] \times \bar{\Omega})$ be the solution operator for the initial value problem (4.16), (4.17). Then, $S_{P,B}$ extends uniquely to a (weak star) continuous map from $CD_{P',B'}^\infty(0)'$ to $\mathcal{D}'([0, T] \times \Omega)$ given by*

$$(4.20) \quad \langle S_{P,B} \Lambda, F \rangle = \langle \Lambda, (\psi(0), \psi_t(0)) \rangle \quad \forall F \in C_0^\infty([0, T] \times \Omega),$$

where ψ is the (smooth) solution of the adjoint problem,

$$P'\psi = F \quad \text{in } [0, T] \times \bar{\Omega}, \quad B'\psi = 0 \quad \text{on } [0, T] \times \partial\Omega,$$

$$\psi(T, \cdot) = \psi_t(T, \cdot) = 0.$$

Proof. If $\Lambda = [(u_0, u_1), \cdot]$ for $(u_0, u_1) \in CD_{P,B}^\infty(0)$, then (4.19) shows that (4.20) is valid for such Λ , where the injection of Proposition 4.3 is implicit. Since such Λ are dense, uniqueness follows. It is easy to verify that (4.20) defines a continuous extension, proving existence. \square

For $(u_0, u_1) \in CD_{P',B'}^\infty(0)'$, the distribution $S_{P,B}(u_0, u_1)$ is called the *generalized solution* of (4.16)–(4.17) with initial data (u_0, u_1) . Generalized solutions of the adjoint problem are defined similarly. Note that generalized solutions satisfy $Pu = 0$ in the sense of $\mathcal{D}'([0, T] \times \Omega)$. Since $[0, T] \times \partial\Omega$ is noncharacteristic for P the traces of the derivatives of u define distributions on $[0, T] \times \partial\Omega$ and satisfy $Bu = 0$ there. The initial conditions (4.17) are valid in the sense of $\mathcal{D}'(\Omega)$. We emphasize, however, that (4.20) is stronger than the union of these conditions.

THEOREM 4.5. *If B is as in (4.12), then the transpose of the map K defined in (4.1)–(4.4) is given by $K'(\Lambda) = \psi|_{\Gamma'}$, where χ is the (generalized) solution of $P'\psi = 0$ in $[0, T] \times \Omega$, $B'\psi = 0$ on $[0, T] \times \partial\Omega$, with Cauchy data $\Lambda \in CD_{P,B}^\infty(T)'$ at time T . If $B = B_{\text{Dir}}$, then the transpose is given by $K'(\Lambda) = \partial_\nu \psi|_{\Gamma'}$.*

Proof. Suppose that B is as in (4.12). Green's identity for $(\psi_0, \psi_1) \in CD_{P',B}^\infty(T)$ and u solving (4.1)–(4.3) yields

$$\int_{]0,T[\times \partial\Omega} \psi g \, dt \, d\sigma = \int_{\Omega} u(T, x) \psi_t(T, x) - u_t(T, x) \psi(T, x) \, dx - \int_{\partial\Omega} au \psi(T, \cdot) \, d\sigma.$$

Thus $K'([\cdot, (\psi_0, \psi_1)]) = \psi|_{\Gamma}$, which is the desired identity. By continuity, the result follows for general functionals $\Lambda \in CD_{P,B}^\infty(T)'$ from its validity on a dense set.

The case where $B = B_{\text{Dir}}$ is similar. \square

For $s \in \mathbb{R}$, we study the range of the operator K when the controls g are chosen in $X \equiv \dot{H}^s(\Gamma)(\dot{H}^{s-1}(\Gamma)$ if $B = B_{\text{Dir}}$) and the hypotheses of Corollary 3.7 are satisfied, that is, P, B is forward and backward well posed, and all compressed null bicharacteristics pass over Γ at nondiffractive points with suitable care taken in case of infinite-order contact. Then the L^2 -wellposedness implies that $Kg = (u(T), u_t(T))$ belongs to $H^s(\Omega) \times H^{s-1}(\Omega)$. It may seem reasonable to try to apply Theorem 4.1 with this choice for the space Y . This is not wise since the data at time t satisfy compatibility conditions, so K is far from onto. For s negative or noninteger, the description of the compatibility conditions is sometimes not obvious. This circle of ideas is the object of the results preceding the exact controllability Theorem 4.9. Once those preliminaries are settled, the theorem follows easily.

For any $s \in \mathbb{R}$ and $k \in \mathbb{N}$, L^2 -wellposedness implies that the map from $CD_{P,B}^\infty(t)$ to $C^\infty([0, T] \times \bar{\Omega})$ is continuous if the domain space is given the $H^s(\Omega) \times H^{s-1}(\Omega)$ topology and the range the topology of $C^k([0, T]; H^{s-k}(\Omega))$. This suggests the following definition.

DEFINITION. $CD_{P,B}^s(t)$ is the closure in $H^s(\Omega) \times H^{s-1}(\Omega)$ of $CD_{P,B}^\infty(t)$. $CD_{P',B'}^s(t)$ is defined similarly.

If $s' > s$, then with each inclusion dense and continuous,

$$(4.21) \quad CD_{P',B'}^\infty(t)' \supset CD_{P,B}^s(t) \supset CD_{P,B}^{s'}(t) \supset CD_{P,B}^\infty(t).$$

Solutions of $Pu = 0$, $Bu = 0$, with initial data in $CD_{P,B}^s(t_1)$ belong to $C^k([0, T]; H^{s-k}(\Omega))$ for all $k \in \mathbb{N}$; thus

$$S_{P,B} : CD_{P,B}^s(0) \rightarrow \cap C^k([0, T]; H^{s-k}(\Omega)).$$

We denote by $S_{P,B}(t_1, t_2)$ the evolution operator from $CD_{P,B}^s(t_1)$ to $CD_{P,B}^s(t_2)$. Similarly, $S_{P',B'}(t_2, t_1)$ maps $CD_{P',B'}^s(t_2)$ to $CD_{P',B'}^s(t_1)$. Hypothesis (3.9) guarantees that $S_{P,B}(t_1, t_2)$ is an isomorphism with inverse $S_{P,B}(t_2, t_1)$.

Green's identity (4.15) yields for $0 \leq t_1 \leq t_2 \leq T$ the transpose relationship

$$[S_{P,B}(t_1, t_2)\Lambda, \Psi] = [\Lambda, S_{P',B'}(t_2, t_1)\Psi]$$

for all $\Lambda \in CD_{P,B}^\infty(t_1)$ and $\Psi \in CD_{P',B'}^\infty(t_2)$. It follows, by continuity, that $S_{P',B'}(t_2, t_1)$ is an isomorphism of $CD_{P,B}^s(t_2)'$ to $CD_{P,B}^s(t_1)'$ equal to the transpose of $S_{P,B}(t_1, t_2)$. The next result lies a little deeper.

PROPOSITION 4.6. *Suppose that P, B is L^2 -well-posed forward and backward in time, $\Lambda \in CD_{P',B'}^\infty(0)'$, and $u = S_{P,B}\Lambda$ is the generalized solution with initial data Λ . Then the following three conditions are equivalent:*

- (i) $\Lambda \in CD_{P,B}^s(0)$;
- (ii) $u \in \cap C^k([0, T]; H^{s-k}(\Omega))$;
- (iii) $u \in H^s(]0, T[\times \Omega)$.

This proposition implies that $CD_{P,B}^s(t)$ is exactly the Cauchy data at time t of $H^s(]0, T[\times \Omega)$ solutions.

Proof. We have already observed that (i) \Rightarrow (ii) \Rightarrow (iii). To prove that (iii) \Rightarrow (i), the key step is the next lemma, which is like Friedrich's classical weak = strong lemma.

LEMMA 4.7. *Suppose that P, B is L^2 -well-posed forward in time, and $u \in H^s([0, T[\times \Omega)$ satisfies $Pu = 0$ in $]0, T[\times \Omega$ and $Bu|_{]0, T[\times \partial\Omega} = 0$. Then there is a sequence $u_n \in C^\infty([0, T] \times \bar{\Omega})$ satisfying (4.16) and $u_n \rightarrow u$ in $H^s([0, T[\times \Omega)$.*

Proof. The key is to mollify u with a smooth approximate delta. The regularization is performed in coordinates that flatten the boundary, and care must be taken that the smoothing kernel has support arranged so that the values of the convolution in $[0, T] \times \bar{\Omega}$ are determined by u in $]0, T[\times \Omega$. The correct strategy is to regularize tangentially first, then in the normal directions. The essential techniques are exposed in [LP], [Sar]. The result is a sequence $v_n \in C^\infty([0, T] \times \bar{\Omega})$ with

$$\begin{aligned} v_n &\rightarrow u && \text{in } H^s([0, T[\times \Omega), \\ Pv_n &\rightarrow 0 && \text{in } H^{s-1}([0, T[\times \Omega), \\ Bv_n &\rightarrow 0 && \text{in } H^{s-1}([0, T[\times \partial\Omega) \quad (H^s \text{ if } B = B_{\text{Dir}}). \end{aligned}$$

Choose $F_n \in C^\infty(\mathbb{R} \times \Omega)$, $g_n \in C^\infty(\mathbb{R} \times \partial\Omega)$ such that $\text{supp } F_n$ and $\text{supp } g_n$ are contained in $\{-1 < t < T+1\}$, $F_n = Pv_n$ and $g_n = Bv_n$ if $0 < t < T$,

$$(4.22) \quad F_n \rightarrow 0 \quad \text{in } H^s(\mathbb{R} \times \Omega),$$

and

$$(4.23) \quad g_n \rightarrow 0 \quad \text{in } H^{s-1}(\mathbb{R} \times \Omega) \quad (H^s \text{ if } B = B_{\text{Dir}}).$$

Let $w_n \in C^\infty(\mathbb{R} \times \bar{\Omega})$ be defined as the solution of the initial value problem

$$Pw_n = -F_n, \quad Bw_n = -g_n, \quad w_n = 0 \quad \text{for } t < -1.$$

Then, thanks to (3.4), (4.22), and (4.23), w_n converges to zero in $H^s([-\infty, T[\times \Omega)$.

The sum $u_n \equiv v_n + w_n$ has the desired properties. \square

Returning to the proof that (iii) \Rightarrow (i), choose u_n as in Proposition 4.6. Then $u_n(0, \cdot)$, $\partial_t u_n(0, \cdot)$ is in $CD_{P,B}^\infty(0)$.

Hypothesis (3.9) guarantees that there is a positive constant c such that for all t_1 and t_2 belonging to $[0, T]$, and all smooth solutions v or (4.16)

$$\|v(t_1, \cdot)\|_{H^s(\Omega)}^2 + \|\partial_t v(t_1, \cdot)\|_{H^{s-1}(\Omega)}^2 < c(\|v(t_2, \cdot)\|_{H^s(\Omega)}^2 + \|\partial_t v(t_2, \cdot)\|_{H^{s-1}(\Omega)}^2).$$

Apply this with $t_1 = 0$ and $v = u_n - u_m$, and integrate dt_2 from 0 to T . The resulting inequality, together with the fact that u_n is a Cauchy sequence in $H^s([0, T[\times \Omega)$, shows that $u_n(0, \cdot)$, $\partial_t u_n(0, \cdot)$ is a Cauchy sequence in $H^s(\Omega) \times H^{s-1}(\Omega)$.

The limit Ψ is an element of $CD_{P,B}^s(0)$, and $S_{P,B}(u_n(0, \cdot), \partial_t u_n(0, \cdot))$ converges in $H^s([0, T[\times \Omega)$ to $S_{P,B}(\Psi)$. However, $S_{P,B}(u_n(0, \cdot), \partial_t u_n(0, \cdot)) = u_n$, which converges to u in the same topology. Thus $S_{P,B}(\Psi) = S_{P,B}(\Lambda)$, which suffices to show that $\Lambda = \Psi$ as elements of $CD_{P,B}^\infty(0)'$. This proves (i). \square

The proof also shows that the norm of u in $H^s([0, T[\times \Omega)$ and the norm of Λ in $CD_{P,B}^s(0)$ are equivalent norms.

For $1 \leq s \in \mathbb{N}$, the analogue of [RM, § 3] shows that $CD_{P,B}^s(t)$ consists exactly of those $H^s(\Omega) \times H^{s-1}(\Omega)$ Cauchy data such that $\partial^j Bu|_{\{t\} \times \partial\Omega} = 0$ for $0 \leq j \leq s-2$ if B is of order one ($0 \leq j \leq s-1$ if $B = B_{\text{Dir}}$). This characterization will not be used below.

We apply Theorem 4.1 with Y equal to CD^s . The next result identifies the dual of CD^s .

PROPOSITION 4.8. *Suppose that P, B satisfies (3.9). Then, for any $s \in \mathbb{R}$, the map $CD_{P,B}^\infty(T) \ni \Psi \mapsto [\cdot, \Psi] \in CD_{P,B}^s(T)'$ extends uniquely to an isomorphism of $CD_{P,B}^{1-s}(T)$ onto $CD_{P,B}^s(T)'$.*

Proof. We prove that there are positive constants c_1 and c_2 , such that, for all $\Psi \in CD_{P',B}^\infty(T)$,

$$(4.24) \quad c_1 \|\Psi\|_{CD_{P',B}^{1-s}(T)} \leq \|[\cdot, \Psi]\|_{CD_{P,B}^s(T)'} \leq c_2 \|\Psi\|_{CD_{P',B}^{1-s}(T)}.$$

It follows that the map of the proposition is an isomorphism onto a closed subspace of $CD_{P,B}^s(T)'$. The density of the image follows from Proposition 4.3.

To prove (4.24), suppose that $\Psi \in CD_{P',B}^\infty(T)$ and let $\psi \equiv S_{P',B}(\Psi)$; so Proposition 4.6 shows that the $CD_{P',B}^{1-s}(T)$ -norm of Ψ is equivalent to the norm of ψ in $H^{1-s}([0, T] \times \Omega)$.

For $F \in C_0^\infty([0, T] \times \Omega)$, let u be the solution of $Pu = F$, $Bu = 0$, with vanishing Cauchy data at $t = 0$. Denote by $U(T)$ the Cauchy data of u at $t = T$. Green's identity shows that $[\Psi, U(T)] = \int \psi F dx dt$. Thus

$$\left| \int \psi F dx dt \right| \leq c_1 \|\Psi\|_{CD_{P',B}^{1-s}(T)} \|U(T)\|_{CD_{P,B}^s(T)} \leq c_2 \|\Psi\|_{CD_{P',B}^{1-s}(T)} \|F\|_{\dot{H}^{s-1}}.$$

This shows that the norm of ψ in $\dot{H}^{s-1}([0, T] \times \Omega)'$ is dominated by a multiple of the $(CD^s)'$ norm of Ψ . Since the $(\dot{H}^{s-1})'$ norm and the $H^{1-s}([0, T] \times \Omega)$ norms are equivalent, this proves the first inequality in (4.24).

For the second inequality, we must estimate $[U, \Psi]$ for arbitrary $U \in CD_{P,B}^s(T)$. Using (3.9), let u be the generalized solution of (4.16) with Cauchy data equal to U at $t = T$. Choose $\chi \in C^\infty([0, T])$ with $\chi = 0$ for $t < T/2$ and χ identically equal to one on $[2T/3, T]$. Then Green's identity applied with u and $\chi(t)\psi$ yields

$$[U, \Psi] = \int u[P, \chi]\psi dx dt, \quad [P, \chi] \equiv \text{commutator of } P \text{ and } \chi.$$

Note that $[P, \chi]$ is a first-order operator involving only time derivatives and with coefficients supported in $[T/2, 2T/3]$. Choose $\eta \in C_0^\infty([0, T])$ with η identically equal to one on the support of the coefficients of $[P, \chi]$.

Introduce local coordinates preserving t and mapping the boundary to $\{x_n = 0\}$. The integral on the right is dominated by the $H_{(0,s)}$ norm of ηu times the $H_{(0,-s)}$ norm of $[P, \chi]\psi$. The latter is dominated by the $H_{(0,1-s)}$ norm of ψ . Theorem B.2.9 of [H] shows that, since $Pu = 0$ and $P'\psi = 0$ and $[0, T] \times \partial\Omega$ is noncharacteristic, these norms are dominated by the $H^s([0, T] \times \Omega)$ -norm of u and the $H^{1-s}([0, T] \times \Omega)$ -norm of ψ , respectively. Thus

$$|[U, \Psi]| < c \|\Psi\|_{CD_{P',B}^{1-s}(T)} \|U\|_{CD_{P,B}^s(T)},$$

which proves the desired estimate for $[\cdot, \Psi]$ in $CD_{P,B}^s(T)'$. \square

THEOREM 4.9. *Suppose that P, B and Γ satisfy the hypotheses of Corollary 3.7. If B is as in (4.12) (respectively, $B = B_{\text{Dir}}$), then for any $s \in \mathbb{R}$, K defined in (4.1)–(4.4) maps $\dot{H}^{s-1}(\Gamma)$ (respectively, $\dot{H}^s(\Gamma)$) onto the annihilator in $CD_{P,B}^s(T)$ of the linear functionals of the form $[\cdot, \Psi]$, where Ψ is the Cauchy datum at time T of an invisible solution, ψ , of $P'\psi = 0$, $B'\psi = 0$.*

Thus the achievable state is an explicitly described finite-codimensional subspace of the H^s solutions of $Pu = 0$, $Bu = 0$ in $t > T$.

Proof. For $g \in C_0^\infty(\Gamma)$, the solution u to (4.1), (4.3) belongs to $C(\mathbb{R}; H^s(\Omega)) \cap C^1(\mathbb{R}; H^{s-1}(\Omega))$, and the map is continuous from $\dot{H}^{s-1}(\Gamma)$ to this space. Thus K extends uniquely to a continuous map, denoted K^s , from $\dot{H}^{s-1}(\Gamma)$ to $CD_{P,B}^s(T)$. The strategy is to apply Theorem 4.1 with $X \equiv \dot{H}^{s-1}(\dot{H}^s \text{ if } B = B_{\text{Dir}})$, $Y \equiv CD_{P,B}^s(T)$, and $X' = H^{-s+1}(\Omega)$ ($H^{-s}(\Omega)$ if $B = B_{\text{Dir}}$).

Thanks to the dense inclusions (4.21), we know that the transpose of K^s is the restriction to $CD_{P,B}^s(T)$ of K' . In particular, the nullspace of $(K^s)'$ is contained in the nullspace of K' .

Since Γ satisfies the hypotheses of Corollary 3.7, and the rays of the adjoint problem are the same as those for the original problem, the conclusions of Corollary 3.7 are also valid for the adjoint problem. The characterization of K' in Theorem 4.5 shows that the kernel of K' coincides with the invisible solutions of the adjoint problem. Corollary 3.7 implies that this is a finite-dimensional subspace of $CD_{P',B'}^\infty(T)$. In particular, the kernels of $(K^s)'$ and K' are identical.

Next, we turn to the verification of (4.5). Suppose that $\Lambda \in Y' = CD_{P,B}^s(T)'$, and let $\psi \equiv S_{P',B'}\Lambda$ be the solution of the adjoint problem with Cauchy data equal to Λ at $t = T$. Then, for B of order one, $\psi|_\Gamma = K'(\Lambda) \in X' = H^{1-s}(\Gamma)$. If $B = B_{\text{Dir}}$, then $\partial_\nu \psi|_\Gamma = K'(\Lambda) \in X' = H^{-s}(\Gamma)$. In either case, Corollary 3.7 implies that $\psi \in H^{1-s}([0, T] \times \Omega)$. In addition, the analogue of (3.10) is

$$(4.25) \quad \|K'(\Lambda)\|_{X'} \geq c_1 \|\psi\|_{H^{-s+1}([0, T] \times \Omega)} - c_2 \|\psi\|_{H^{-s}([0, T] \times \Omega)}.$$

The remark following Proposition 4.6 shows that the norms on the right are equivalent to the norms of Λ in $CD_{P',B'}^{-s+1}(T)$ and $CD_{P',B'}^{-s}(T)$, respectively.

Proposition 4.8 shows that the $CD_{P',B'}^{-s+1}(T)$ -norm is equivalent to the Y' norm.

Finally, Rellich's compactness theorem shows that $H^{-s+1}([0, T] \times \Omega)$ is compactly included in $H^{-s}([0, T] \times \Omega)$.

The last three assertions show that (4.25) is an inequality of the form (4.5) with X and Y chosen as above.

Then Theorem 4.1 together with the identification of the nullspace of K^s given above proves the desired result. \square

COROLLARY 4.10. *If the hypotheses of Theorem 3.8 are satisfied, then $\mathcal{J} = \{0\}$ and the set of achievable states for controls in $\dot{H}^{s-1}(\Gamma)$ ($\dot{H}^s(\Gamma)$ if $B = B_{\text{Dir}}$) is exactly the space $CD_{P,B}^s(T)$.*

Thus, for time-independent problems, the rest state can be steered to an arbitrary H^s solution.

Examples. Suppose that $P = \square$, $B = B_{\text{Dir}}$ or $B = \partial/\partial\nu$, and $\Gamma =]0, T[\times \omega$ has the property that every generalized bicharacteristic passes over a nondiffractive point in Γ . Then the boundary conditions are well posed in both directions, so Corollary 4.10 applies. We describe the results obtained upon taking $s = 0, 1, 2$ in the following cases.

1. When $B = B_{\text{Dir}}$, $CD^1 = \dot{H}^1(\Omega) \times L^2(\Omega)$ and $CD^0 = L^2(\Omega) \times \dot{H}^{-1}$. These are the achievable states for controls in $\dot{H}^1(\Gamma)$ and $L^2(\Gamma)$, respectively. For $s = 2$, $CD^2 = (H^2(\Omega) \cap \dot{H}^1(\Omega)) \times \dot{H}^1(\Omega)$ and the controls are in $\dot{H}^2(\Gamma)$.

2. When $B = \partial/\partial\nu$, $CD^1 = H^1(\Omega) \times L^2(\Omega)$, and $CD^0 = L^2(\Omega) \times H^{-1}(\Omega)$. These are the achievable states for controls in $L^2(\Gamma)$ and $\dot{H}^{-1}(\Gamma)$, respectively. For $s = 2$, $CD^2 = \{u \in H^2(\Omega) : \partial_\nu u|_{\partial\Omega} = 0\} \times H^1(\Omega)$ and the controls are in $\dot{H}^1(\Gamma)$.

As mentioned after Theorem 3.8, the result $\mathcal{J} = \{0\}$ for time-dependent P, B can sometimes be proved using the Holmgren uniqueness theorem.

5. Stabilization of waves. The problem of stabilization is related to but different from the problem of exact controllability. In the problem of controllability, we are asked to steer a solution from a known initial condition to a desired final condition by action on Γ , determined from the known initial and final state. In the problem of stabilization, we do not know the state, but hope, by intervention at Γ , to damp the motion to zero. The intervention at $x \in \Gamma$ is determined locally by the values of the derivatives of u at x . We imagine a machine that reacts to the local state. A mathematical formulation is to suppose that, away from Γ , we have a boundary condition representing

free motion, while on Γ there is a different condition whose aim is to damp the waves. One then wants conditions which guarantee that solutions tend to zero as t tends to infinity. What is more, we would like the rate of convergence to be uniform on bounded sets of data.

We can often show that stabilizability implies controllability. This has been one way that exact controllability results have been proved. This section discusses stabilization in its own right. The exact controllability results that are consequences do not add to the information from § 3. Attention is called to one important difference. For controllability, it is important that the boundary conditions be reversible in the sense that (3.9) is satisfied. This is so that efforts to control the waves are not lost when the waves undergo subsequent reflections. For stabilization, such reversibility is not needed, and, in fact, the best stabilization is often achieved by irreversible boundary conditions.

Example. Consider $u_t - u_{xx} = 0$ on $]0, 1[$ with boundary conditions

$$u_t - u_x|_{x=0} = 0 \quad \text{and} \quad u_t + u_x|_{x=1} = 0.$$

Then, for any initial data, the solution is identically zero when $t \geq 1$. This is completely efficient stabilization by radically irreversible boundary conditions.

For the reversible conditions

$$u_t - (1 - \varepsilon)u_x|_{x=0} = 0 \quad \text{and} \quad u_t + (1 - \varepsilon)u_x|_{x=1} = 0, \quad 0 < \varepsilon \ll 1,$$

the evolution operator for one unit of time, $S(1)$, has norm of order ε . The irreversible choice is best possible here.

We get good results for problems that are dissipative and time independent. More precisely, suppose that

$$(5.1) \quad P = \partial_t^2 - \sum \partial_i a_{ij}(x) \partial_j + c(x), \quad c \text{ real-valued},$$

and the boundary condition $Bu = 0$ satisfies

$$(5.2) \quad \text{on each component of } \partial\Omega \text{ either } B = B_{\text{Dir}} \text{ or } B = \partial_\nu + \alpha(x)\partial_t + \beta(x),$$

where α, β are smooth real-valued functions on $\partial\Omega$. A class of nonlocal conditions with links to elasticity and to absorbing artificial boundary conditions in numerical analysis is discussed in [BHLRZ]. To simplify the expressions for the energy and the law of energy decay, we set $\alpha = \beta = 0$ on those components of the boundary where $B = B_{\text{Dir}}$. We further suppose that Ω is connected.

Smooth solutions of $Pu = 0$, $Bu = 0$ then satisfy the energy law

$$(5.3) \quad \partial_t e = \int_{\partial\Omega} \alpha(x) |u_t|^2 d\sigma,$$

where the energy e is defined as

$$(5.4) \quad 2e(t) \equiv \int_{\Omega} |u_t|^2 + a_{ij} \partial_i u \partial_j \bar{u} + c(x) |u|^2 dx + \int_{\partial\Omega} \beta(x) |u|^2 d\sigma.$$

We first discuss the case where the energy is positive definite in the sense that

There is a $c > 0$ such that

$$(5.5) \quad \int_{\Omega} a_{ij} \partial_i v \partial_j \bar{v} + c(x) |v|^2 dx + \int_{\partial\Omega} \beta(x) |v|^2 d\sigma \geq c \|v\|_{H^1(\Omega)}^2$$

for all $v \in H^1$ that vanish on the components of $\partial\Omega$,
where $B = B_{\text{Dir}}$.

This is guaranteed if $c \geq 0$, $\beta \geq 0$ and, at least one of them is not identically zero or $B = B_{\text{Dir}}$. On at least one component of $\partial\Omega$. At the end of the section, we will discuss the case where $c \equiv 0$, $\beta \equiv 0$ for which (5.5) is violated.

For time-independent operators, $CD_{P,B}^s(t)$ (defined in § 3) is independent of time and is abbreviated as $CD_{P,B}^s$. We omit the P, B when there is little chance for confusion. Hypothesis (5.5) implies that, for $U = (u_0, u_1) \in CD^\infty$,

$$(5.6) \quad \|(u_0, u_1)\|^2 \equiv \int_{\Omega} a_{ij} \partial_i u_0 \partial_j \bar{u}_0 + c(x) |v_0|^2 dx + \int_{\partial\Omega} \beta(x) |v_0|^2 d\sigma + \|u_1\|_{L^2}^2$$

is a norm equivalent to the norm in $H^1(\Omega) \times L^2(\Omega)$, which is the norm in $CD_{P,B}^1 = \{v \in H^1(\Omega): \alpha v|_{\partial\Omega} = 0\} \times L^2(\Omega)$. If $\alpha \geq 0$ and CD^1 is normed by (5.6), then the evolution operator on CD^∞ is CD^1 -norm-decreasing. Thus the evolution operators, $S_{P,B}(0, t)$ define a contraction semigroup on CD^1 . When there is little risk of confusion, we simply write $S(t)$.

THEOREM 5.1. *Suppose that hypotheses (5.1), (5.2), and (5.5) hold, α is not identically zero, and $CD_{P,B}^1$ is normed by (5.6). Then $S(t)$ defines a contraction semigroup on CD^1 , and, as t tends to $+\infty$, $s\text{-}\lim S(t) = 0$ in $CD_{P,B}^1$.*

The strong limit means that, for any $U \in CD^1$, $\|S(t)U\|_{CD^1} \rightarrow 0$ as $t \rightarrow +\infty$. Thus the energy of finite energy solutions decays to zero as t tends to infinity. This can be proved using Iwasaki's [I] criterion as in [BLR3] or using the argument of Haraux [Ha1] as in [BLR1, Prop. 8].

If u is a smooth solution of $Pu = 0$, $Bu = 0$ in $\{t \geq 0\} \times \bar{\Omega}$, then the same is true of $\partial_t u$ and, more generally, $\partial_t^j u$ for any $j \geq 1$. Hypothesis (5.5) guarantees that for $1 \leq s \in \mathbb{N}$, the norm

$$(5.7) \quad \sum_{1 \leq j \leq s} \|(\partial_t^j u(t, \cdot), \partial_t^{j-1} u(t, \cdot))\|_{CD_{P,B}^1},$$

where CD^1 is normed by (5.6), is equivalent to the norm in $H^s \times H^{s-1}$, which is the norm in CD^s .

COROLLARY 5.2. *Suppose that the hypotheses of Theorem 5.1 hold, $1 \leq s \in \mathbb{N}$, and that $CD_{P,B}^s$ is normed by (5.7). Then the evolution operator $S(t)$ defines a contraction semigroup on CD^s , and, for any $U \in CD^s$, $\|S(t)U\|_{CD^s} \rightarrow 0$ as $t \rightarrow +\infty$.*

Remark. This result can be extended to nonpositive $0 \geq s \in \mathbb{Z}$ using the duality in Proposition 4.8, the fact that $S_{P,B}(0, t)' = S_{P',B'}(0, -t)$, and the fact that $S_{P',B'}(0, -t)$ tends strongly to zero on $CD_{P',B'}^{1-s}$ thanks to Corollary 5.2.

The problem of stabilization is to determine under what conditions the decay of $S(t)$ is uniform on bounded subsets of CD^s . Thanks to the semigroup property of S and the close relation between S on CD^1 and S on CD^s there are many equivalent formulations.

PROPOSITION 5.3. *Assume the hypotheses of Corollary 5.2 hold. Then the following are equivalent:*

1. *As t tends to infinity, $S(t)U$ tends to zero in CD^1 uniformly for U in bounded subsets of CD^1 ;*
2. *There is a $T > 0$ such that the norm of $S(T)$ in $\text{Hom}(CD^1)$ is less than one;*
3. *There are constants M and $\omega > 0$ such that $\|S(t)\|_{\text{Hom}(CD^1)} \leq M e^{-\omega t}$ for all $t \geq 0$;*
4. *For all $s \in \mathbb{Z}$, conditions 1–3 hold with CD^1 replaced by CD^s .*

The existence of finite-dimensional spaces of arbitrarily high dimension that are localized near rays as in Theorem 3.2 shows that the geometric conditions that played a role in observation and control are also crucial for stabilization.

THEOREM 5.4. *If $\Gamma \equiv]0, T[\times \{\alpha > 0\}$ and there is a generalized bicharacteristic γ that does not pass over Γ and has at most finite-order contact with $T^*(\mathbb{R} \times \partial\Omega)$, then for*

$s \in \mathbb{Z}$, $\|S_{P,B}(T)\|_{\text{Hom}(CD^s)} = 1$. Even more, the intersection of the essential spectrum of $S_{P,B}(T)$ with the unit circle $\{|z| = 1\}$ is nonempty.

In particular, to have stabilization, it is necessary that there exist a $T > 0$ so that every generalized bicharacteristic with at most finite-order contact with $T^*(\mathbb{R} \times \partial\Omega)$ must pass over $\{\alpha > 0\}$.

This necessary condition is not far from sufficient. For sufficiency, we suppose that compressed bicharacteristics pass through nondiffractive points in $T^*(\Gamma)$.

THEOREM 5.5. *Suppose that the hypotheses of Corollary 5.2 hold and that there is a $T > 0$ such that $\Gamma \equiv]0, T[\times \{\alpha > 0\}$ satisfies the hypotheses of either Theorem 3.3 or Theorem 3.4. Then, for all $s \in \mathbb{Z}$, $\|S(T)\|_{\text{Hom}(CD^s)} < 1$. In particular, there is stabilization.*

Proof. The general case where $z \in \mathbb{Z}$ follows from the case where $s = 1$. For any $\eta > 0$, let $\Gamma_\eta \equiv]\eta, T - \eta[\times \{\alpha > \eta\}$. A compactness argument shows that, for η sufficiently small, Γ_η satisfies the hypotheses of either Theorem 3.3 or Theorem 3.4.

For a finite energy solution u , use (3.8) applied to the set Γ_η to get a lower bound for the energy dissipated for $0 < t < T$. From (5.3) the dissipation is given by

$$(5.8) \quad e(0) - e(T) = \int \int_{\Gamma} \alpha(x) |u_t|^2 dt d\sigma \geq (\eta/2) \int \int_{\Gamma_{\eta/2}} \alpha(x) |u_t|^2 dt d\sigma.$$

Inequality (3.8) implies that

$$(5.9) \quad \|\partial_\nu u\|_{L^2(\Gamma_\eta)} + \|u\|_{H^1(\Gamma_\eta)} \geq c \|u\|_{H^1(]T-\varepsilon, T[\times \Omega)} - c' \|u\|_{L^2(]0, T[\times \Omega)}$$

Since S is contractive in CD^1 , we have

$$(5.10) \quad \|u\|_{H^1(]T-\varepsilon, T[\times \Omega)} \geq (\varepsilon)^{1/2} \|u(T), u_t(T)\|_{CD^1}.$$

Proposition 4.6 implies that

$$(5.11) \quad \|u\|_{L^2(]0, T[\times \Omega)} \leq c \|u(0), u_t(0)\|_{CD^0}.$$

Since $\text{WF}_b(u) \cap T^*(\partial M)$ is contained in $\mathcal{H} \cup \mathcal{G}$, on which ∂_t is elliptic, and $\partial_\nu u + \alpha \partial_t u + \beta u = 0$ on Γ , there are constants $c, c' > 0$ such that

$$(5.12) \quad \|u_t\|_{L^2(\Gamma_{\eta/2})} \geq c \|\partial_\nu u\|_{L^2(\Gamma_\eta)} + c \|u\|_{H^1(\Gamma_\eta)} - c' \|u\|_{L^2(\Gamma_{\eta/2})}.$$

Since $\alpha > \eta$ on Γ_η ,

$$(5.13) \quad \int \int_{\Gamma} \alpha(x) |u_t|^2 dt d\sigma \geq c \|u_t\|_{L^2(\Gamma_{\eta/2})}^2.$$

Finally, using [Ho, Thm. B.2.9] as in the proof of (3.15), yields

$$(5.14) \quad \|u\|_{L^2(\Gamma_{\eta/2})} \leq c \|u\|_{H^{1/2}(]0, T[\times \Omega)} \leq c' \|u(0), u_t(0)\|_{CD^{1/2}}.$$

Combining inequalities (5.9)–(5.14), there is a $c > 0$ and c' such that all finite energy solutions satisfy

$$(5.15) \quad \int \int_{\Gamma} \alpha(x) |u_t|^2 dt d\sigma \geq c \|u(T), u_t(T)\|_{CD^1}^2 - c' \|u(0), u_t(0)\|_{CD^{1/2}}^2.$$

If the last term on the right were absent, this, together with (5.9), would imply that $\|S(T)\| \leq (1 + c)^{-1} < 1$. Thus it remains to prove that there is a $c'' > 0$ such that

$$(5.16) \quad \int \int_{\Gamma} \alpha(x) |u_t|^2 dt d\sigma \geq c'' \|u(T), u_t(T)\|_{CD^1}^2.$$

If this were not the case, there would exist a sequence of CD^1 solutions such that

$$(5.17) \quad \|u_n(T), \partial_t u_n(T)\|_{CD^1} = 1 \quad \text{and} \quad \int_{\Gamma} \int_0^T \alpha |\partial_t u_n|^2 dt d\sigma \rightarrow 0.$$

From (5.15) we see that

$$(5.18) \quad \liminf \|u_n(0), \partial_t u_n(0)\|_{CD^{1/2}} \geq 1/c' > 0.$$

From (5.8) we have

$$\lim \|u_n(0), \partial_t u_n(0)\|_{CD^1} = 1.$$

Since CD^1 is compactly imbedded in $CD^{1/2}$, we may pass to a subsequence, still denoted u_n , such that

$$(u_n(0), \partial_t u_n(0)) \rightarrow (u(0), \partial_t u(0)) \quad \text{in } CD^{1/2}.$$

$$(u_n(0), \partial_t u_n(0)) \rightharpoonup (u(0), \partial_t u(0)) \quad \text{in } CD^1.$$

By (5.18) the limit is nonzero. From (5.17) we conclude that the solution with Cauchy data equal to $u(0), \partial_t u(0)$ satisfies $\partial_t u|_{\Gamma} = 0$. Thus, to complete the proof of (5.16) and therefore the proof of the theorem, it suffices to show that the only $H^1([0, T[\times \Omega)$ solution of $Pu = 0$, $Bu = 0$, which satisfies $\partial_t u|_{\Gamma} = 0$ is $u = 0$.

Denote by \mathcal{N} the set of $u \in H^{1/2}([0, T[\times \Omega)$, which satisfy $Pu = 0$, $Bu = 0$, and, $\partial_t u|_{\Gamma} = 0$. For $u \in \mathcal{N}$, (5.12) and (5.14) yield

$$\|\partial_\nu u\|_{L^2(\Gamma_\eta)} + \|u\|_{H^1(\Gamma_\eta)} \leq c \|u(0), \partial_t u(0)\|_{CD^{1/2}}.$$

This, together with (3.8) applied to Γ_η with η small, yields

$$\|u(0), \partial_t u(0)\|_{CD^1} \leq c \|u(0), \partial_t u(0)\|_{CD^{1/2}}.$$

Since CD^1 is compactly imbedded in $CD^{1/2}$, Reisz's theorem implies that $\dim(\mathcal{N}) < \infty$. Repeating the above bootstrap starting with the fact that $u(0), \partial_t u(0) \in CD^1$ yields $u(0), \partial_t u(0) \in CD^{3/2}$. Thus $\partial_t u$ is then an $H^{1/2}$ solution satisfying $Pu = 0$, $Bu = 0$, $u_t|_{\Gamma} = 0$, that is, $\partial_t u \in \mathcal{N}$.

Thus ∂_t is a linear map of the finite-dimensional space \mathcal{N} to itself. If \mathcal{N} were nonempty, there would be a nonzero eigenfunction u , $\partial_t u = \lambda u$. Then $u = e^{\lambda t} v(x)$. The equation $Pu = 0$ implies that

$$(5.19) \quad -\sum \partial_i a_{ij}(x) \partial_j v + (\lambda^2 + c(x))v = 0.$$

If $\lambda \neq 0$, the fact that $u_t = 0$ on Γ implies that $v = 0$ on Γ . Then the boundary condition $(\partial_\nu + \alpha(x)\lambda + \beta)v = 0$ implies that $\partial_\nu v = 0$ in Γ . Then uniqueness in the Cauchy problem for (5.19) implies that $v \equiv 0$.

If $\lambda = 0$, multiply (5.19) by \bar{v} and integrate over Ω to find that

$$\int_{\Omega} a_{ij} \partial_i v \partial_j \bar{v} + c(x) |v|^2 dx + \int_{\partial\Omega} \beta(x) |v|^2 d\sigma = 0.$$

Hypothesis (5.5) implies that $v = 0$. Thus there can be no nonzero eigenfunctions, so $\mathcal{N} = \{0\}$. \square

A unique continuation argument like the end of the proof is a crucial ingredient in the proof of Theorem 5.1. In fact, the continuation argument can be replaced by Theorem 5.1 as follows. The decay to zero of the solution $u = e^{\lambda t} v(x)$ shows that $\operatorname{Re} \lambda < 0$. Thus $e(1) < e(0)$, so u_t is not identically zero on Γ , a contradiction.

An interesting case where (5.5) is violated is when $c(x) \equiv 0$, $\beta \equiv 0$, and the boundary operator is nowhere equal to B_{Dir} . Then constants are solutions of $Pu = 0$, $Bu = 0$, and expression (5.6) vanishes when v is a constant function. It is reasonable to expect that the adjoint problem also has a stationary solution and that applying this linear functional to an arbitrary solution u of $Pu = 0$, $Bu = 0$ should yield a nontrivial conservation law. The conservation law is that

$$\int_{\Omega} u_t dx + \int_{\partial\Omega} \alpha(x) u d\sigma = \text{independent of } t.$$

The verification is simply by differentiating with respect to time and using the differential equation and boundary conditions. Replacing u by $u - (\int u_t(0, \cdot) dx - \int \alpha u(0, \cdot) d\sigma) / \int \alpha d\sigma$ reduces us to the study of solutions for which

$$\int_{\Omega} u_t dx + \int_{\partial\Omega} \alpha(x) u d\sigma = 0.$$

Denote by $CD^{s,0}$ the set of u_0, u_1 in CD^s such that

$$\int_{\Omega} u_1(x) dx + \int_{\partial\Omega} \alpha(x) u_0 d\sigma = 0.$$

Then $CD^{1,0}$ is normed by expression (5.6), and then (5.7) norms $CD^{s,0}$. A proof exactly like that of Theorem 5.4 yields the following result.

THEOREM 5.6. *Suppose that (5.1) holds with $c(x) \equiv 0$, $B = \partial_\nu + \alpha(x)\partial_t$, $\alpha \geq 0$, and $\Gamma \equiv]0, T[\times \{\alpha > 0\}$ satisfies the hypotheses of either Theorem 3.3 or Theorem 3.4. For $1 \leq s \in \mathbb{N}$, norm $CD^{s,0}$ as above. Then $\|S(T)\| < 1$ as a map of $CD^{s,0}$ to itself. In particular, for bounded sets of data in CD^s , $u - (\int u_t(0, \cdot) dx - \int \alpha u(0, \cdot) d\sigma) / \int \alpha d\sigma$ converges to zero in CD^s at a uniform exponential rate.*

Acknowledgments. In the preparation of this work, the authors have benefited from many suggestions from colleagues. In particular, the authors thank C. Borgers, P. Gerard, I. Lasieka, J. L. Lions, R. Melrose, O. Pironeau, M. Taylor, R. Triggiani, and M. Williams for their valuable advice.

REFERENCES

- [A] J. A. ARNAUD, *Hamiltonian theory of beam propagation*, in Progress in Optics XI, E. Wolf, ed., North-Holland, Amsterdam, 1973, pp. 249–304.
- [AM] K. ANDERSSON AND R. MELROSE, *The propagation of singularities along gliding rays*, Invent. Math., 41 (1977), pp. 197–232.
- [BHLRZ] C. BARDOS, L. HALPERN, G. LEBEAU, J. RAUCH, AND E. ZUAZUA, *Stabilization de l'équation des ondes au moyen d'un feedback portant sur la condition aux limites de Dirichlet*, Asymptotic Anal., to appear.
- [BLR1] C. BARDOS, G. LEBEAU, AND J. RAUCH, *Contrôle et stabilisation dans les problèmes hyperboliques*, in Controlabilité Exacte Perturbations et Stabilisation de Système Distribués, Appendice II, Tome 1, Collection RMA, Masson, Paris, 1988.
- [BLR2] ———, *Un exemple d'utilisation des notions de propagation pour le contrôle et la stabilisation des problèmes hyperboliques*, Rend. Sem. Mat. Univ. Pol. Torino, Fasc. Spec. 1988, pp. 11–32.
- [BLR3] ———, *Microlocal ideas in control and stabilization*, Proc. of Clermont-Ferrand Colloquium, Springer Lecture Notes in Control & Information Science, 125 (1989), pp. 1–30, Springer-Verlag, Berlin, New York, 1988.
- [C] J. CHAZARAIN, *Construction de la paramétrix du problème mixte hyperbolique pour l'équation des ondes*, C. R. Acad. Sci. Paris, 276 (1973), pp. 1213–1215.
- [CP] J. CHAZARAIN AND A. PIRIOU, *Introduction à la Théorie des Equations aux Dérivées Partielles Linéaires*, Gauthiers-Villars, Paris, 1981.

- [G] L. GARDING, *Le problème de la dérivée oblique pour l'équation des ondes*, C. R. Acad. Sci. Paris, 285 (1977), pp. 773–775; rectification, 285 (1978), p. 1199.
- [GLL] R. GLOWINSKI, C. LI, AND J. L. LIONS, to appear.
- [Ha1] A. HARAUX, *Stabilisation of trajectories for some weakly damped hyperbolic equations*, J. Differential Equations, 59 (1985), pp. 145–154.
- [Ha2] ———, *Contrôlabilité exacte d'une membrane rectangulaire au moyen d'une fonctionnelle analytique localisée*, C. R. Acad. Sci. Paris, 1988.
- [Ho] L. HORMANDER, *The Analysis of Linear Partial Differential Operators*, Vols. I, III, Springer-Verlag, Berlin, Heidelberg, 1983, 1985.
- [I] N. IWASAKI, *Local decay of solutions for symmetric hyperbolic systems with dissipative and coercive boundary conditions in exterior domains*, Publ. RIMS Kyoto, 5 (1969), pp. 193–218.
- [LP] P. D. LAX AND R. S. PHILLIPS, *Local boundary conditions for dissipative symmetric linear differential operators*, Comm. Pure Appl. Math., 13 (1960), pp. 427–455.
- [Le1] G. LEBEAU, *Contrôle et stabilisation hyperboliques*, Séminaire Equations aux Dérivées Partielles, Ecole Polytechnique, Paris, 1989–90.
- [Le2] ———, *Contrôle de l'équation de Schrödinger*, J. Analyse Math., to appear.
- [LM] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
- [Lio] J. L. LIONS, *Contrôlabilité Exacte Perturbations et Stabilisation de Systèmes Distribués*, Tome 1, Collection RMA, Masson, Paris, 1988.
- [Lit1] W. LITTMAN, *Boundary control theory for hyperbolic and parabolic partial differential equations with constant coefficients*, Ann. Scuola Nor. Sup. Pisa, V (1978), pp. 567–580.
- [Lit2] ———, *Near optimal time boundary controllability for a class of hyperbolic equations*, in Proc. Conference on Distributed Parameter Control, Gainesville, GA, Springer-Verlag, New York, 1986.
- [Me] R. MELROSE, *Transformation of boundary problems*, Acta Math., 147 (1981), pp. 149–236.
- [MS1] R. MELROSE AND J. SJOSTRAND, *Singularities of boundary value problems I*, Comm. Pure Appl. Math., 31 (1978), pp. 593–617.
- [MS2] ———, *Singularities of boundary value problems, II*, Comm. Pure Appl. Math., 35 (1982), pp. 129–168.
- [MRS] C. MORAWETZ, J. RALSTON, AND W. STRAUSS, *Decay of solutions of the wave equation outside a nontrapping obstacle*, Comm. Pure Appl. Math., 30 (1977), pp. 447–508.
- [Mi] S. MIYATAKE, *Mixed problems for hyperbolic equations of second order with first order complex boundary operators*, Japan J. Math., 1975, pp. 111–158.
- [Na] O. NALIN, *Contrôlabilité exacte sur une partie du bord des équations de Maxwell*, C. R. Acad. Sci. Paris, Sér. 1, 309 (1989), pp. 811–815.
- [N] L. NIRENBERG, *Lectures on Linear Partial Differential Equations*, Regional Conference Series in Math. #17, American Mathematical Society, Providence, RI, 1967.
- [Ra1] J. RALSTON, *Solutions of the wave equation with localised energy*, Comm. Pure Appl. Math., 22 (1969), pp. 807–823.
- [Ra2] ———, *Gaussian beams and the propagation of singularities*, in Studies in Partial Differential Equations, W. Littman, ed., MAA Studies in Math., Vol. 23, 1982, pp. 206–248.
- [RM] J. RAUCH AND F. MASSEY, *Differentiability of solutions to hyperbolic initial-boundary value problems*, Trans. Amer. Math. Soc., 189 (1974), pp. 303–318.
- [RS] J. RAUCH AND J. SJOSTRAND, *Propagation of analytic singularities along diffracted rays*, Indiana Univ. Math. J., 30 (1981), pp. 389–401.
- [RT1] J. RAUCH AND M. TAYLOR, *Exponential decay of solutions to hyperbolic equations in bounded domains*, Indiana Univ. Math. J., 24 (1974).
- [RT2] ———, *Penetration into shadow regions and unique continuation properties in hyperbolic mixed problems*, Indiana Univ. Math. J., 22 (1972), pp. 277–285.
- [Ru] D. L. RUSSELL, *Controllability and stabilization theory for linear partial differential equations. Recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.
- [Sak] R. SAKAMOTO, *Hyperbolic Boundary Value Problems*, K. Miyahara, trans., Cambridge University Press, Cambridge, UK, 1982.
- [Sar] L. SARASON, *On weak and strong solutions of boundary value problems*, Comm. Pure Appl. Math., 15 (1962), pp. 237–288.
- [Sj] J. SJOSTRAND, *Propagation of analytic singularities for second-order Dirichlet problems*, Comm. Partial Differential Equations, 1980, pp. 41–94.
- [T1] M. TAYLOR, *Reflection of singularities of solutions to systems of differential equations*, Comm. Pure Appl. Math., 28 (1975), pp. 457–478.
- [T2] ———, *Pseudodifferential Operators*, Princeton University Press, Princeton, NJ, 1981.

EXACT BOUNDARY CONTROLLABILITY OF THE WAVE EQUATION AS THE LIMIT OF INTERNAL CONTROLLABILITY*

CAROLINE FABRE†

Abstract. This paper presents the study of the following problem of exact controllability concerning the wave equation with Dirichlet boundary conditions. Using Lions's Hilbert uniqueness method (HUM), Zuazua has given a positive answer to the problem of exact controllability when the control is distributed and acts on an ε -neighborhood of a part Γ_0 of the boundary satisfying some geometrical conditions. The main interest is in the passage to the limit when ε goes to 0, which means when the neighborhood of Γ_0 shrinks to Γ_0 itself.

Key words. exact internal controllability, boundary controllability, singular perturbations

AMS(MOS) subject classifications. 35B25, 49E25, 93B05

1. Introduction. We present here the study of the following problem of exact controllability concerning the wave equation with Dirichlet boundary conditions. Using Lions's Hilbert uniqueness method (HUM) Zuazua [8] has given a positive answer to the problem of exact controllability when the control is distributed and acts on an ε -neighborhood of a part Γ_0 of the boundary satisfying some geometrical conditions. We are interested here in the passage to the limit when ε goes to 0, which means when the neighborhood of Γ_0 shrinks to Γ_0 itself.

To solve this problem, we must study the convergence of solutions of the wave equation with homogeneous boundary Dirichlet conditions and with singular right-hand sides concentrated in an ε -neighborhood of a part Γ_0 of the boundary, and more precisely the convergence (when ε goes to 0) of

$$\begin{aligned}\psi_\varepsilon'' - \Delta \psi_\varepsilon &= \tilde{\varphi}_\varepsilon \chi_{\omega_\varepsilon \times (0, T)} \quad \text{in } Q, \\ \psi_\varepsilon &= 0 \quad \text{on } \Sigma, \\ \psi_\varepsilon(0) &= y^0 \quad \text{and} \quad \psi_\varepsilon'(0) = y^1, \\ \psi_\varepsilon(T) &= 0 \quad \text{and} \quad \psi_\varepsilon'(T) = 0,\end{aligned}$$

where ω_ε is a neighborhood of Γ_0 , $\chi_{\omega_\varepsilon \times (0, T)}$ is the characteristic function of $\omega_\varepsilon \times (0, T)$, and $(y^0, y^1) \in H_0^1(\Omega) \times L^2(\Omega)$. Note that the right-hand side is concentrated in a neighborhood of Γ_0 and is singular because we will prove that $\|\tilde{\varphi}_\varepsilon \chi_{\omega_\varepsilon \times (0, T)}\|_{L^2(\omega_\varepsilon \times (0, T))} = O(\varepsilon^{-3/2})$.

We will see that under the conditions that are naturally given by our controllability problem, we obtain, when ε tends to 0, a solution of the homogeneous wave equation but with a nonhomogeneous Dirichlet boundary condition.

We prove that the limit problem of the internal exact controllability is the result given by HUM for the problem of exact controllability when the control acts on the Dirichlet boundary condition on Γ_0 . For this purpose, we will use all the results that are announced in [4] and proved in [5] concerning the behavior near the boundary for solutions of the wave equation.

Some analogous questions may be considered for other types of equations, and we can refer to [3] for the study of the Schrödinger equation and the application to

* Received by the editors December 26, 1990; accepted for publication June 14, 1991.

† Ecole Polytechnique, Centre de Mathématiques Appliquées, Unité de Recherche Associée au Centre National de la Recherche Scientifique (CNRS)-756, 91128 Palaiseau Cedex, France.

the dynamic plate equation, and to [2], where the case of the beams equation is considered.

In § 2, we briefly recall the results obtained by Zuazua about exact controllability of the wave equation when the control acts on an ε -neighborhood of Γ_0 and belongs to L^2 . These results can be found in [8]. In § 3, we establish the estimate on the controls in terms of ε . The passage to the limit when ε goes to 0 is then studied in § 4. We finish in § 5 by the similar problem when we change the space of initial data.

2. Exact internal controllability of the wave equation. Let Ω be a bounded open set with a C^3 -boundary Γ , and let $\nu(y)$ be the unit exterior normal at a point y of Γ . Let Γ_0 be an open subset of Γ . We consider the following condition on Γ_0 , which has been introduced by Lions in [6] or [7]:

$$(2.1) \quad \exists x_0 \in \mathbb{R}^N \quad \text{such that } \Gamma_0 = \{x \in \Gamma \text{ such that } (x - x_0) \cdot \nu(x) > 0\}.$$

We write

$$\omega_\varepsilon = \bigcup_{x \in \Gamma_0} (B(x, \varepsilon) \cap \Omega),$$

where $B(x, \varepsilon)$ denotes the open ball centered at x with radius ε . For $T > 0$, we write $\Sigma_0 = \Gamma_0 \times (0, T)$, $Q = \Omega \times (0, T)$, $\Sigma = \Gamma \times (0, T)$, $Q_\varepsilon = \omega_\varepsilon \times (0, T)$, and $\chi_{\omega_\varepsilon \times (0, T)}$ the characteristic function of Q_ε .

The problem solved by Zuazua is the following theorem.

THEOREM 2.1. *If Γ_0 satisfies (2.1), there exists $T_0 > 0$ such that for any $\varepsilon > 0$ and $T > T_0$ and any (y^0, y^1) in $H_0^1(\Omega) \times L^2(\Omega)$, there exists a control $v_\varepsilon \in L^2(\omega_\varepsilon \times (0, T))$ such that the solution ψ_ε of*

$$(2.2) \quad \begin{aligned} \psi_\varepsilon'' - \Delta \psi_\varepsilon &= \begin{cases} v_\varepsilon & \text{in } Q_\varepsilon, \\ 0 & \text{in } Q - Q_\varepsilon, \end{cases} \\ \psi_\varepsilon &= 0 \quad \text{on } \Sigma, \\ \psi_\varepsilon(0) &= y^0 \quad \text{and} \quad \psi_\varepsilon'(0) = y^1 \end{aligned}$$

satisfies $\psi_\varepsilon(T) = \psi_\varepsilon'(T) = 0$.

Remark 2.1. The time T_0 is equal to $2R(x_0)$, where $R(x_0)$ is the smallest radius of the ball centered at x_0 that contains Ω .

To establish this theorem, Zuazua uses Lions's HUM, which leads us to define v_ε as a solution of the homogeneous wave equation in the following way.

For $(\varphi^0, \varphi^1) \in L^2(\Omega) \times H^{-1}(\Omega)$, we consider the (weak) solution φ of the homogeneous wave equation

$$(2.3) \quad \begin{aligned} \varphi'' - \Delta \varphi &= 0 \quad \text{in } Q, \\ \varphi &= 0 \quad \text{on } \Sigma, \\ \varphi(0) &= \varphi^0 \quad \text{and} \quad \varphi'(0) = \varphi^1. \end{aligned}$$

From φ , we define ψ as the solution of

$$(2.4) \quad \begin{aligned} \psi'' - \Delta \psi &= \varphi \chi_{\omega_\varepsilon \times (0, T)} \quad \text{in } Q, \\ \psi &= 0 \quad \text{on } \Sigma, \\ \psi(T) &= 0 \quad \text{and} \quad \psi'(T) = 0. \end{aligned}$$

We know from the regularity results concerning the solutions of the wave equation (refer to [7]) that

$$(2.5) \quad \text{the mapping } (\varphi^0, \varphi^1) \in L^2(\Omega) \times H^{-1}(\Omega) \rightarrow \psi \in C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega)) \text{ is linear continuous.}$$

We then define the operator Λ_ε from $L^2(\Omega) \times H^{-1}(\Omega)$ into $L^2(\Omega) \times H_0^1(\Omega)$ by

$$(2.6) \quad \Lambda_\varepsilon(\varphi^0, \varphi^1) = (-\psi'(0), \psi(0)).$$

From (2.5), Λ_ε is continuous, and ψ will be a solution of (2.2) if and only if $\Lambda_\varepsilon(\varphi^0, \varphi^1) = (-y^1, y^0)$. Hence, we are led to prove that Λ_ε is invertible. For this, we multiply (2.4) by φ , and we obtain

$$(2.7) \quad -(\varphi^0, \psi'(0)) + \langle \varphi^1, \psi(0) \rangle = \int_0^T \int_{\omega_\varepsilon} \varphi^2(x, t) \, dx \, dt,$$

where (\cdot) denotes the scalar product in $L^2(\Omega)$, and $\langle \cdot \rangle$ the duality $H^{-1}(\Omega)$, $H_0^1(\Omega)$.

The following theorem, proved by Zuazua, states the invertibility of Λ_ε .

THEOREM 2.2. *If Γ_0 satisfies (2.1), there exists $c(\varepsilon) > 0$ such that for any (φ^0, φ^1) in $L^2(\Omega) \times H^{-1}(\Omega)$, the solution φ of (2.3) satisfies*

$$(2.8) \quad \|\varphi^0\|_{L^2(\Omega)}^2 + \|\varphi^1\|_{H^{-1}(\Omega)}^2 \leq c(\varepsilon) \int_0^T \int_{\omega_\varepsilon} \varphi^2(x, t) \, dx \, dt.$$

Remark 2.2. At this stage, we have no information on the estimate of $c(\varepsilon)$ in terms of ε .

Now, for (y^0, y^1) fixed in $H_0^1(\Omega) \times L^2(\Omega)$, define $\tilde{\varphi}_\varepsilon^0 \in L^2(\Omega)$ and $\tilde{\varphi}_\varepsilon^1 \in H^{-1}(\Omega)$ by $(\tilde{\varphi}_\varepsilon^0, \tilde{\varphi}_\varepsilon^1) = \Lambda_\varepsilon^{-1}(-y^1, y^0)$. Then, we define ψ_ε by (2.4) taking as right-hand side the solution $\tilde{\varphi}_\varepsilon$ of (2.3) satisfying $\tilde{\varphi}_\varepsilon(0) = \tilde{\varphi}_\varepsilon^0$ and $\tilde{\varphi}_\varepsilon'(0) = \tilde{\varphi}_\varepsilon^1$. From the definition of Λ_ε , we have $\psi_\varepsilon(0) = y^0$ and $\psi_\varepsilon'(0) = y^1$; hence ψ_ε is also a solution of (2.2) and satisfies $\psi_\varepsilon(T) = \psi_\varepsilon'(T) = 0$.

To study the passage to the limit when ε goes to 0 in the problems of exact controllability given by Theorem 2.1, we first need estimates on the controls v_ε . To get them, we are going to precise the dependence of the constant $c(\varepsilon)$ of Theorem 2.2 in terms of ε .

3. Estimate on $c(\varepsilon)$. Before estimating the constant $c(\varepsilon)$, we must study a preliminary result concerning the convergence of some linear forms that will appear naturally later on. We begin with a result that we will use in this section and the following one.

THEOREM 3.1. *Let Γ_0 be an open subset of Γ_0 . Let $(\varphi_\varepsilon^0)_\varepsilon$ and $(\varphi_\varepsilon^1)_\varepsilon$ be two sequences weakly converging in L^2 and $H^{-1}(\Omega)$. We suppose that there exist $C > 0$ and $\varepsilon_0 > 0$ such that the solutions φ_ε of (2.3) where $\varphi_\varepsilon(0) = \varphi_\varepsilon^0$ and $\varphi_\varepsilon'(0) = \varphi_\varepsilon^1$ satisfy*

$$\frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} \varphi_\varepsilon^2(x, t) \, dx \, dt \leq C \quad \forall \varepsilon \in]0, \varepsilon_0[.$$

We introduce the linear forms

$$(3.1) \quad \begin{aligned} G_\varepsilon : H_0^1(\Omega) \times L^2(\Omega) \times L^1(0, T; L^2(\Omega)) &\rightarrow \mathbb{R}, \\ (u^0, u^1, h) &\rightarrow \frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} \varphi_\varepsilon(x, t) u(x, t) \, dx \, dt, \end{aligned}$$

where u is the solution of

$$(3.2) \quad \begin{aligned} u'' - \Delta u &= h \quad \text{in } Q, \\ u &= 0 \quad \text{on } \Sigma, \\ u(0) &= u^0 \quad \text{and} \quad u'(0) = u^1. \end{aligned}$$

Then

$$(3.3) \quad \text{the limit } \varphi \text{ of } (\varphi_\varepsilon)_\varepsilon \text{ (in } L^\infty(0, T; L^2(\Omega)) \text{ weak-*) satisfies } \frac{\partial \varphi}{\partial \nu} \in L^2(\Sigma_0),$$

and the linear forms G_ε are bounded in $H^{-1}(\Omega) \times L^2(\Omega) \times L^\infty(0, T; L^2(\Omega))$ after extraction of a subsequence, they converge for the weak-* topology of this space to

$$(3.4) \quad \begin{aligned} G : H_0^1(\Omega) \times L^2(\Omega) \times L^1(0, T; L^2(\Omega)) &\rightarrow \mathbb{R}, \\ (u^0, u^1, h) &\rightarrow \frac{1}{3} \int_{\Sigma_0} \frac{\partial \varphi}{\partial \nu}(y, t) \frac{\partial u}{\partial \nu}(y, t) dy dt. \end{aligned}$$

Furthermore,

$$(3.5) \quad \frac{1}{3} \int_{\Sigma_0} \left| \frac{\partial \varphi}{\partial \nu}(y, t) \right|^2 dy dt \leq \liminf_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} \varphi_\varepsilon^2(x, t) dx dt.$$

Remark 3.1. If Γ_0 satisfies (2.1), from Lions's result (see [6] or [7]) saying that for $T > T_0$, the $L^2(\Sigma_0)$ -norm of the normal derivative on Σ_0 is equivalent to the $H_0^1(\Omega) \times L^2(\Omega)$ norm of the initial data for solutions of (2.3), we deduce from (3.3) that the limit φ has a finite energy.

Proof of Theorem 3.1. The proofs of (3.3) and (3.5) are given in [5]. We first prove that the sequence $(G_\varepsilon)_\varepsilon$ is bounded in $H^{-1}(\Omega) \times L^2(\Omega) \times L^\infty(0, T; L^2(\Omega))$. To describe its limit, we will proceed in several steps. From Hölder's inequality, we get

$$\begin{aligned} |G_\varepsilon(u^0, u^1, h)|^2 &\leq \left(\frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} \varphi_\varepsilon^2(x, t) dx dt \right) \left(\frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} u^2(x, t) dx dt \right), \\ &\leq C \left(\frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} u^2(x, t) dx dt \right). \end{aligned}$$

In [5], we have proved the following theorem.

THEOREM 3.2. *There exists $c > 0$ independent on ε such that for every (u^0, u^1, h) in $H_0^1(\Omega) \times L^2(\Omega) \times L^1(0, T; L^2(\Omega))$, the solution u of (3.2) satisfies*

$$(3.6) \quad \frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} u^2(x, t) dx dt \leq c(\|h\|_{L^1(0, T; L^2(\Omega))}^2 + \|u^0\|_{H_0^1(\Omega)}^2 + \|u^1\|_{L^2(\Omega)}^2).$$

Using Theorem 3.2, we have

$$|G_\varepsilon(u^0, u^1, h)| \leq \sqrt{Cc}(\|h\|_{L^1(0, T; L^2(\Omega))} + \|u^0\|_{H_0^1(\Omega)} + \|u^1\|_{L^2(\Omega)}),$$

and the forms G_ε are bounded in $H^{-1}(\Omega) \times L^2(\Omega) \times L^\infty(0, T; L^2(\Omega))$. So there exists G in $H^{-1}(\Omega) \times L^2(\Omega) \times L^\infty(0, T; L^2(\Omega))$ such that (after extraction of a subsequence) G_ε converges for the weak-* topology of $H^{-1}(\Omega) \times L^2(\Omega) \times L^\infty(0, T; L^2(\Omega))$ to G . To find G , we introduce the following linear forms:

$$\begin{aligned} L_\varepsilon : L^2(0, T; H^2(\Omega) \cap H_0^1(\Omega)) &\rightarrow \mathbb{R}, \\ v &\rightarrow \frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} \varphi_\varepsilon(x, t) v(x, t) dx dt, \end{aligned}$$

and we prove Lemma 3.1.

LEMMA 3.1. *Under the hypotheses of Theorem 3.1, the linear forms L_ε defined above are bounded in $L^2(0, T; (H^2(\Omega) \cap H_0^1(\Omega)))'$, and (after extraction of a subsequence) they converge for the weak topology of this space to*

$$L: L^2(0, T; H^2(\Omega) \cap H_0^1(\Omega)) \rightarrow \mathbb{R},$$

$$v \rightarrow \frac{1}{3} \int_{\Sigma_0} \frac{\partial \varphi}{\partial \nu}(y, t) \frac{\partial v}{\partial \nu}(y, t) dy dt.$$

Proof of Lemma 3.1. We first show that $(L_\varepsilon)_\varepsilon$ is bounded in $L^2(0, T; (H^2 \cap H_0^1(\Omega)))'$. For $v \in L^2(0, T; H^2 \cap H_0^1(\Omega))$, we have

$$|L_\varepsilon(v)| \leq \sqrt{C} \left(\frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} |v(x, t)|^2 \right)^{1/2}.$$

We can prove (see, for example, [6]) that there exists a covering of Γ consisting in open sets U_1, \dots, U_p satisfying the following lemma.

LEMMA 3.2. *There exist open sets U_1, \dots, U_p , and $\varepsilon_0 > 0$ such that*

$$(3.7) \quad \overline{\omega_\varepsilon} \subset \bigcup_{i=1}^p U_i, \quad \text{where } \overline{\omega_\varepsilon} \text{ denotes the closure of } \omega_\varepsilon \quad \forall \varepsilon \in]0, \varepsilon_0];$$

$$(3.8) \quad \exists ! (y, z) \in (\Gamma \cap U_i) \times \mathbb{R}_+ \text{ such that } x = y - z\nu(y) \quad \forall x \in \Omega \cap U_i,$$

(we will note $y = p(x)$);

$$(3.9) \quad \text{The mappings } J_i^{-1}: x \rightarrow (y, z) \text{ are } C^2\text{-diffeomorphisms from } \Omega \cap U_i \text{ on their images, which map } \omega_\varepsilon \cap U_i \text{ into } (\Gamma \cap U_i) \times]0, \varepsilon[;$$

$$(3.10) \quad \text{There exist } \varepsilon_0 > 0, m > 0, \text{ and } M > 0, \text{ such that}$$

$$m \leq |J_i(y, z)| \leq M \quad \forall z \in [0, \varepsilon_0], \forall i,$$

where $|J_i(y, z)|$ denotes the Jacobian of J_i at the point (y, z) ,

$$(3.11) \quad |J_i(y, z)| = 1 \quad \forall y \in \Gamma \cap U,$$

$$(3.12) \quad \text{The mappings } (y, z) \rightarrow |J_i(y, z)| \text{ are } C^1.$$

Finally, if v is a function defined on $\Omega \cap U$, we note that $\hat{v}(y, z) = v(x)$, and we have, if $v \in H^1(\Omega \cap U)$,

$$(3.13) \quad \frac{\partial \hat{v}}{\partial z}(y, z) = -\nabla v(x) \cdot \nu(p(x)), \quad \text{where } p(x) = y.$$

Let $\alpha_1, \dots, \alpha_p$ be a partition of unity relative to U_1, \dots, U_p . Writing $v_i = \alpha_i v$, we have

$$(3.14) \quad \frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} |v(x, t)|^2 \leq 2^{p-1} \sum_{k=1}^p \frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon \cap U_k} |\hat{v}_k(y, z, t)|^2 |J_k(y, z)| dy dz dt.$$

However,

$$(3.15) \quad \hat{v}_k(y, z, t) = z \frac{\partial \hat{v}_k}{\partial z}(y, 0, t) + \int_0^z \int_0^s \frac{\partial^2 \hat{v}_k}{\partial z^2}(y, r, t) dr ds;$$

hence, putting (3.15) in (3.14) and taking into account the properties relative to J_k of Lemma 3.2, we get

$$\frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} |v(x, t)|^2 \leq c \|v\|_{L^2(0, T; H^2 \cap H_0^1)},$$

where c depends only on Ω and T .

The sequence $(L_\varepsilon)_\varepsilon$ is now bounded in $L^2(0, T; (H^2 \cap H_0^1(\Omega)))'$. Let L be its limit after extraction of a subsequence and for the weak topology of $L^2(0, T; (H^2 \cap H_0^1(\Omega)))'$. First, we determine L on a dense subspace of $L^2(0, T; H^2 \cap H_0^1(\Omega))$ by Lemma 3.3.

LEMMA 3.3. *For $v \in D([0, T]; H^2 \cap H_0^1(\Omega))$, we have*

$$\lim_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} \varphi_\varepsilon(x, t) v(x, t) \, dx \, dt = \frac{1}{3} \int_0^T \int_{\Gamma_0} \frac{\partial \varphi}{\partial \nu} \frac{\partial v}{\partial \nu} \, dy \, dt.$$

Proof of Lemma 3.3. This proof is quite long and technical.

We introduce

$$\tilde{\omega}_\varepsilon = \{x \in \omega_\varepsilon \text{ such that } \exists (y, z) \in \Gamma_0 \times]0, \varepsilon[, x = y - zv(y)\};$$

so $\tilde{\omega}_\varepsilon \subseteq \omega_\varepsilon$, and

$$(3.16) \quad m[(\partial \omega_\varepsilon - \partial \tilde{\omega}_\varepsilon) \cap \Gamma] \xrightarrow{\varepsilon \rightarrow 0} 0,$$

where m denotes the boundary measure.

For $v \in D([0, T]; H^2 \cap H_0^1(\Omega))$, we write $L_\varepsilon(v)$ as

$$(3.17) \quad \begin{aligned} L_\varepsilon(v) &= \frac{1}{\varepsilon^3} \int_0^T \int_{\tilde{\omega}_\varepsilon} \varphi_\varepsilon(x, t) v(x, t) \, dx \, dt + \frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon - \tilde{\omega}_\varepsilon} \varphi_\varepsilon(x, t) v(x, t) \, dx \, dt \\ &= A_\varepsilon(v) + B_\varepsilon(v), \end{aligned}$$

and we study separately $A_\varepsilon(v)$ and $B_\varepsilon(v)$.

Let us first show that $B_\varepsilon(v)$ goes to 0. Since

$$|B_\varepsilon(v)| \leq \sqrt{C} \left(\frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon - \tilde{\omega}_\varepsilon} |v(x, t)|^2 \, dx \, dt \right)^{1/2},$$

it is sufficient to prove that

$$\frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon - \tilde{\omega}_\varepsilon} |v(x, t)|^2 \, dx \, dt \xrightarrow{\varepsilon \rightarrow 0} 0.$$

Using Lemma 3.2 and (3.14), we get

$$\begin{aligned} & \frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon - \tilde{\omega}_\varepsilon} |v(x, t)|^2 \, dx \, dt \\ & \leq 2^{p-1} \sum_{k=1}^p \frac{1}{\varepsilon^3} \int_0^T \int_{\partial(\omega_\varepsilon - \tilde{\omega}_\varepsilon) \cap \Gamma \cap U_k} \int_0^\varepsilon |\hat{v}_k(y, z, t)|^2 |J_k(y, z)| \, dy \, dz \, dt, \\ & \leq 2^p \sum_{k=1}^p \left(\frac{M_k}{3} \int_0^T \int_{\partial(\omega_\varepsilon - \tilde{\omega}_\varepsilon) \cap \Gamma \cap U_k} \left| \frac{\partial \hat{v}_j}{\partial z}(y, 0, t) \right|^2 \, dy \, dt \right. \\ & \quad \left. + \frac{1}{\varepsilon^3} \int_0^T \int_{\partial(\omega_\varepsilon - \tilde{\omega}_\varepsilon) \cap \Gamma \cap U_k} \int_0^\varepsilon \left(\int_0^z \int_0^s \left| \frac{\partial^2 \hat{v}_k}{\partial z^2}(y, r, t) \right| \, dr \, ds \right)^2 \right. \\ & \quad \left. \cdot |J_k(y, z)| \, dy \, dz \, dt \right). \end{aligned}$$

We then write

$$\left(\int_0^z \int_0^s \left| \frac{\partial^2 \hat{v}_k}{\partial z^2}(y, r, t) \right| dr ds \right)^2 \leq z \int_0^z \left(s \int_0^s \left| \frac{\partial^2 \hat{v}_k}{\partial z^2}(y, r, t) \right|^2 dr \right) ds,$$

which allows us to easily prove that there exists $c > 0$ independent on ε such that

$$\begin{aligned} & \frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon - \tilde{\omega}_\varepsilon} |v(x, t)|^2 dx dt \\ & \leq c \left(\int_0^T \int_{\partial(\omega_\varepsilon - \tilde{\omega}_\varepsilon) \cap \Gamma} \left| \frac{\partial v}{\partial \nu}(y, t) \right|^2 dy dt + \varepsilon \|v\|_{L^2(0, T; H^2 \cap H_0^1)}^2 \right). \end{aligned}$$

From (3.16), we conclude that $B_\varepsilon(v) \rightarrow 0$ when $\varepsilon \rightarrow 0$. Let us now study $A_\varepsilon(v) = 1/\varepsilon^3 \int_0^T \int_{\tilde{\omega}_\varepsilon} \varphi_\varepsilon(x, t) v(x, t) dx dt$. We have

$$A_\varepsilon(v) = \sum_{j,k=1}^p A_{j,k,\varepsilon}(v),$$

with

$$A_{j,k,\varepsilon}(v) = \frac{1}{\varepsilon^3} \int_0^T \int_{\tilde{\omega}_\varepsilon} \varphi_{j,\varepsilon}(x, t) v_k(x, t) dx dt.$$

We write $\gamma_{j,k} = \partial(\tilde{\omega}_\varepsilon \cap U_j \cap U_k) \cap \Gamma$, $\Sigma_{j,k} = \gamma_{j,k} \times (0, T)$. Using Lemma 3.2, it follows that

$$A_{j,k,\varepsilon}(v) = \frac{1}{\varepsilon^3} \int_0^T \int_{\gamma_{j,k}} \int_0^\varepsilon \hat{\varphi}_{j,\varepsilon}(y, z, t) \hat{v}_k(y, z, t) |J_k(y, z)| dy dz dt.$$

Since $\hat{v}_k \in L^2(0, T; H^2(\tilde{\omega}_\varepsilon \cap U_k))$ and since $L^2(0, T; H^2(\tilde{\omega}_\varepsilon \cap U_k))$ is continuously imbedded in $C_z^1([0, \delta]; L^2((0, T) \times \gamma_{j,k}))$, the following identity in $L^2((0, T) \times \gamma_{j,k})$, (since $\hat{v}_k(y, 0, t) = 0$) holds:

$$(3.18) \quad \hat{v}_k(z) = z \frac{\partial \hat{v}_k}{\partial z}(0) + z V_k(z) \quad \text{with} \quad \lim_{z \rightarrow 0} V_k(z) = 0.$$

Thus

$$\begin{aligned} A_{j,k,\varepsilon}(v) &= \frac{1}{\varepsilon^3} \int_0^T \int_{\gamma_{j,k}} \int_0^\varepsilon \hat{\varphi}_{j,\varepsilon}(y, z, t) z \frac{\partial \hat{v}_k}{\partial z}(y, 0, t) |J_k(y, z)| dy dz dt \\ &+ \frac{1}{\varepsilon^3} \int_0^T \int_{\gamma_{j,k}} \int_0^\varepsilon \hat{\varphi}_{j,\varepsilon}(y, z, t) z V_k(y, z, t) |J_k(y, z)| dy dz dt. \end{aligned}$$

We write

$$C_{j,k,\varepsilon}(v) = \frac{1}{\varepsilon^3} \int_0^T \int_{\gamma_{j,k}} \int_0^\varepsilon \hat{\varphi}_{j,\varepsilon}(y, z, t) z \frac{\partial \hat{v}_k}{\partial z}(y, 0, t) |J_k(y, z)| dy dz dt$$

and

$$D_{j,k,\varepsilon}(v) = \frac{1}{\varepsilon^3} \int_0^T \int_{\gamma_{j,k}} \int_0^\varepsilon \hat{\varphi}_{j,\varepsilon}(y, z, t) z V_k(y, z, t) |J_k(y, z)| dy dz dt.$$

We have

$$|D_{j,k,\varepsilon}(v)| \leq M_k \frac{1}{\varepsilon^3} \int_0^\varepsilon z \|\hat{\varphi}_{j,\varepsilon}(z)\|_{L^2(\Sigma_{j,k})} \|V_k(z)\|_{L^2(\Sigma_{j,k})} dz,$$

so

$$|D_{j,k,\varepsilon}(v)| \leq M_k \frac{1}{\varepsilon^3} \left(\int_0^\varepsilon \|\hat{\phi}_{j,\varepsilon}(z)\|_{L^2(\Sigma_{j,k})}^2 dz \right)^{1/2} \left(\int_0^\varepsilon z^2 \|V_k(z)\|_{L^2(\Sigma_{j,k})}^2 dz \right)^{1/2}.$$

By hypothesis, $1/\varepsilon^3 \int_0^\varepsilon \|\hat{\phi}_{j,\varepsilon}(z)\|_{L^2(\Sigma_{j,k})}^2 dz$ is uniformly bounded in ε ; hence there exists $c > 0$ independent of ε such that

$$|D_{j,k,\varepsilon}(v)| \leq c \|V_k\|_{L_z^\infty(0,\varepsilon;L^2(\Sigma_{j,k}))}.$$

Now, using (3.18), this proves that

$$(3.19) \quad \lim_{\varepsilon \rightarrow 0} D_{j,k,\varepsilon}(v) = 0.$$

We must now study $C_{j,k,\varepsilon}(v)$. For this purpose, we introduce

$$\phi_\varepsilon(t) = \int_0^t \varphi_\varepsilon(\tau) d\tau + \theta_\varepsilon \quad \text{with} \quad \Delta \theta_\varepsilon = \varphi_\varepsilon^1, \theta_\varepsilon \in H_0^1(\Omega).$$

The function ϕ_ε is solution of (2.3) with $\phi_\varepsilon(0) = \theta_\varepsilon$ and $\phi'_\varepsilon(0) = \varphi_\varepsilon^0$. Since the sequence $(\varphi_\varepsilon^1)_\varepsilon$ weakly converges in $H^{-1}(\Omega)$,

$$(3.20) \quad (\theta_\varepsilon)_\varepsilon \text{ weakly converges in } H_0^1(\Omega).$$

We then write that

$$(3.21) \quad |J_k(y, z)| = |J_k(y, 0)| + \int_0^z \frac{\partial}{\partial z} (|J_k(y, s)|) ds = 1 + \int_0^z \frac{\partial}{\partial z} (|J_k(y, s)|) ds.$$

Putting (3.21) in $C_{j,k,\varepsilon}(v)$ and integrating by parts in time, we obtain the following since \hat{v}_k has a compact support in time:

$$\begin{aligned} C_{j,k,\varepsilon}(v) &= -\frac{1}{\varepsilon^3} \int_0^T \int_{\gamma_{j,k}} \int_0^\varepsilon \hat{\phi}_{j,\varepsilon}(y, z, t) z \frac{\partial \hat{v}_{k'}}{\partial z}(y, 0, t) dy dz dt \\ &\quad - \frac{1}{\varepsilon^3} \int_0^T \int_{\gamma_{j,k}} \int_0^\varepsilon \hat{\phi}_{j,\varepsilon}(y, z, t) z \frac{\partial \hat{v}_k'}{\partial z}(y, 0, t) \\ &\quad \cdot \left(\int_0^z \frac{\partial}{\partial z} (|J_k(y, s)|) ds \right) dy dz dt \\ &= E_{j,k,\varepsilon}(v) + F_{j,k,\varepsilon}(v). \end{aligned}$$

From Lemma 3.2, $\int_0^z (\partial/\partial z) (|J_k(y, s)|) ds = O(z)$, uniformly in y . On the other hand,

$$(3.22) \quad \hat{\phi}_{j,\varepsilon}(y, z, t) = \int_0^z \frac{\partial \hat{\phi}_{j,\varepsilon}}{\partial z}(y, r, t) dr;$$

hence there exists $c > 0$ independent of ε such that

$$\begin{aligned} |F_{j,k,\varepsilon}(v)| &\leq c \frac{1}{\varepsilon^3} \left(\int_0^T \int_{\gamma_{j,k}} \int_0^\varepsilon \left(\int_0^z \frac{\partial \hat{\phi}_{j,\varepsilon}}{\partial z}(y, r, t) dr \right)^2 dy dz dt \right)^{1/2} \\ &\quad \cdot \left(\int_0^T \int_{\gamma_{j,k}} \int_0^\varepsilon z^4 \left| \frac{\partial \hat{v}_k'}{\partial z}(y, 0, t) \right|^2 dy dz dt \right)^{1/2}. \end{aligned}$$

We can check easily now that

$$|F_{j,k,\varepsilon}(v)| \leq c\sqrt{\varepsilon} \left\| \frac{\partial \hat{\phi}_{j,\varepsilon}}{\partial z} \right\|_{L^2(\tilde{\omega}_\varepsilon \cap U_j \cap U_k)} \left\| \frac{\partial \hat{v}_k'}{\partial z}(0) \right\|_{L^2(\Sigma_{j,k})}.$$

From (3.20), $(\phi_\varepsilon)_\varepsilon$ is bounded in $C([0, T]; H_0^1(\Omega))$, thus $(\partial \hat{\phi}_{j,\varepsilon} / \partial z)_\varepsilon$ is bounded in $L^2(\tilde{\omega}_\varepsilon \cap U_j \cap U_k)$, and

$$\lim_{\varepsilon \rightarrow 0} F_{j,k,\varepsilon}(v) = 0.$$

Now, consider

$$E_{j,k,\varepsilon}(v) = -\frac{1}{\varepsilon^3} \int_0^T \int_{\gamma_{j,k}} \int_0^\varepsilon \hat{\phi}_{j,\varepsilon}(y, z, t) z \frac{\partial \hat{v}'_k}{\partial z}(y, 0, t) dy dz dt.$$

From (3.22), we get

$$\begin{aligned} E_{j,k,\varepsilon}(v) &= -\frac{1}{\varepsilon^3} \int_0^\varepsilon z \int_0^z \left\langle \frac{\partial \hat{\phi}_{j,\varepsilon}}{\partial z}(r) - \frac{\partial \hat{\phi}_{j,\varepsilon}}{\partial z}(0); \frac{\partial \hat{v}'_k}{\partial z}(0) \right\rangle_{L^2(\Sigma_{j,k})} dr dz \\ &\quad - \frac{1}{3} \left\langle \frac{\partial \hat{\phi}_{j,\varepsilon}}{\partial z}(0); \frac{\partial \hat{v}'_k}{\partial z}(0) \right\rangle_{L^2(\Sigma_{j,k})}. \end{aligned}$$

Let us prove that $(\varphi_\varepsilon)_\varepsilon$ is bounded in $H^{-2}(0, T; H^2(\Omega) \cap H_0^1(\Omega))$. For this, we introduce

$$\psi_\varepsilon(x, t) = \int_0^t \phi_\varepsilon(x, s) ds + \psi_\varepsilon^0 \quad \text{with} \quad \Delta \psi_\varepsilon^0 = \phi_\varepsilon^1 = \varphi_\varepsilon^0, \quad \psi_\varepsilon^0 \in H^2 \cap H_0^1(\Omega).$$

The function ψ_ε is solution of (2.3) and $\psi_\varepsilon'' = \varphi_\varepsilon$. On another hand, since v has a compact support in time,

$$\begin{aligned} E_{j,k,\varepsilon}(v) &= -\frac{1}{\varepsilon^3} \int_0^\varepsilon z \int_0^z \left\langle \frac{\partial \hat{\psi}_{j,\varepsilon}}{\partial z}(r) - \frac{\partial \hat{\psi}_{j,\varepsilon}}{\partial z}(0); \frac{\partial \hat{v}''_k}{\partial z}(0) \right\rangle_{L^2(\Sigma_{j,k})} dr dz \\ &\quad - \frac{1}{3} \left\langle \frac{\partial \hat{\psi}_{j,\varepsilon}}{\partial z}(0); \frac{\partial \hat{v}''_k}{\partial z}(0) \right\rangle_{L^2(\Sigma_{j,k})}. \end{aligned}$$

As $(\varphi_\varepsilon^0)_\varepsilon$ is bounded in $L^2(\Omega)$, $(\psi_\varepsilon^0)_\varepsilon$ is bounded in $H^2 \cap H_0^1(\Omega)$ and $(\psi_\varepsilon)_\varepsilon$ is bounded in $C([0, T]; H^2 \cap H_0^1(\Omega))$. The sequence $(\varphi_\varepsilon)_\varepsilon$ is then bounded in $H^{-2}(0, T; H^2(\Omega) \cap H_0^1(\Omega))$ and we can easily prove that this implies that the functions

$$r \rightarrow \left\langle \frac{\partial \hat{\psi}_{j,\varepsilon}}{\partial z}(r) - \frac{\partial \hat{\psi}_{j,\varepsilon}}{\partial z}(0); \frac{\partial \hat{v}''_k}{\partial z}(0) \right\rangle,$$

are continuous at 0, uniformly in ε . We then deduce that

$$\frac{1}{\varepsilon^3} \int_0^\varepsilon z \int_0^z \left\langle \frac{\partial \hat{\psi}_{j,\varepsilon}}{\partial z}(r) - \frac{\partial \hat{\psi}_{j,\varepsilon}}{\partial z}(0); \frac{\partial \hat{v}''_k}{\partial z}(0) \right\rangle_{L^2(\Sigma_{j,k})} dr dz \xrightarrow{\varepsilon \rightarrow 0} 0.$$

Hence

$$\begin{aligned} \lim_{\varepsilon \rightarrow 0} L_\varepsilon(v) &= \lim_{\varepsilon \rightarrow 0} A_\varepsilon(v) = \lim_{\varepsilon \rightarrow 0} \sum_{j,k=1}^p E_{j,k,\varepsilon}(v) \\ &= -\lim_{\varepsilon \rightarrow 0} \sum_{j,k=1}^p \frac{1}{3} \left\langle \frac{\partial \hat{\phi}_{j,\varepsilon}}{\partial z}(0); \frac{\partial \hat{v}'_k}{\partial z}(0) \right\rangle_{L^2(\Sigma_{j,k})} \\ &= -\lim_{\varepsilon \rightarrow 0} \frac{1}{3} \left\langle \sum_{j=1}^p \frac{\partial \hat{\phi}_{j,\varepsilon}}{\partial z}(0); \sum_{k=1}^p \frac{\partial \hat{v}'_k}{\partial z}(0) \right\rangle_{L^2(\Sigma_0)} \\ &= -\lim_{\varepsilon \rightarrow 0} \frac{1}{3} \left\langle \frac{\partial \phi_\varepsilon}{\partial \nu}; \frac{\partial v'}{\partial \nu} \right\rangle_{L^2(\Sigma_0)}. \end{aligned}$$

As

$$\left\langle \frac{\partial \phi_\varepsilon}{\partial \nu}, \frac{\partial v'}{\partial \nu} \right\rangle_{L^2(\Sigma_0)} = \left\langle \frac{\partial \phi_\varepsilon}{\partial \nu}, \frac{\partial v'}{\partial \nu} \right\rangle_{D'(\Sigma_0), D(\Sigma_0)}$$

and as $(\partial \phi_\varepsilon / \partial \nu)_\varepsilon$ weakly converges to $\partial \phi / \partial \nu$ in $H^{-1}(0, T; L^2(\Gamma_0))$, we have

$$\lim_{\varepsilon \rightarrow 0} L_\varepsilon(v) = \frac{1}{3} \left\langle \frac{\partial \phi}{\partial \nu}, \frac{\partial v}{\partial \nu} \right\rangle_{H^{-1}(0, T; L^2(\Gamma_0)); H_0^1(0, T; L^2(\Gamma_0))}.$$

Furthermore, we know that, at the limit, $\partial \phi / \partial \nu \in L^2((0, T) \times \Gamma_0)$, which gives a sense to the quantity

$$\frac{1}{3} \int_0^T \int_{\Gamma_0} \frac{\partial \phi}{\partial \nu} \frac{\partial v}{\partial \nu} dy dt,$$

and this finishes the proof of Lemma 3.3.

Proof of Lemma 3.1. Let v be in $L^2(0, T; H^2 \cap H_0^1(\Omega))$. Since $D(0, T; H^2 \cap H_0^1(\Omega))$ is dense in $L^2(0, T; H^2 \cap H_0^1(\Omega))$, we can construct a sequence $(v_n)_n$ of elements of $D(0, T; H^2 \cap H_0^1(\Omega))$ such that $v_n \rightarrow v$ in $L^2(0, T; H^2 \cap H_0^1(\Omega))$.

Now, write

$$L_\varepsilon(v) - L(v) = L_\varepsilon(v - v_n) + L_\varepsilon(v_n) - L(v_n) + L(v - v_n).$$

As $\partial \phi / \partial \nu \in L^2((0, T) \times \Gamma_0)$, L is a linear continuous form on $L^2(0, T; H^2 \cap H_0^1(\Omega))$. Furthermore, using the boundedness of $(L_\varepsilon)_\varepsilon$ in $L^2(0, T; (H^2 \cap H_0^1(\Omega))')$, we can easily prove that $L_\varepsilon(v)$ tends to $L(v)$ when ε goes to 0.

Proof of Theorem 3.1. For $(u^0, u^1, h) \in H_0^1(\Omega) \times L^2(\Omega) \times L^1(0, T; L^2(\Omega))$ we consider the solution u of (3.2) associated to this data. We introduce sequences $(v_n^0)_n, (v_n^1)_n$, and $(h_n)_n$ in $H^2 \cap H_0^1(\Omega)$, $H_0^1(\Omega)$, and $L^1(0, T; H_0^1(\Omega))$ such that

$$v_n^0 \rightarrow u^0 \text{ in } H_0^1(\Omega), \quad v_n^1 \rightarrow u^1 \text{ in } L^2(\Omega),$$

and

$$h_n \rightarrow h \text{ in } L^1(0, T; H_0^1(\Omega)).$$

Let v_n be the solution of (3.2) associated to the data v_n^0, v_n^1 , and h_n .

In particular, we have $v_n \in L^2(0, T; H^2 \cap H_0^1(\Omega))$. As $G_\varepsilon(v_n^0, v_n^1, h_n) = L_\varepsilon(v_n)$,

$$\lim_{\varepsilon \rightarrow 0} G_\varepsilon(v_n^0, v_n^1, h_n) = L(v_n) = G(v_n^0, v_n^1, h_n).$$

The linear form defined by (3.12), is continuous on $H_0^1(\Omega) \times L^2(\Omega) \times L^1(0, T; L^2(\Omega))$. Writing then

$$\begin{aligned} G_\varepsilon(u^0, u^1, h) - G(u^0, u^1, h) &= G_\varepsilon[(u^0, u^1, h) - (v_n^0, v_n^1, h_n)] \\ &\quad + [G_\varepsilon(v_n^0, v_n^1, h_n) - G(v_n^0, v_n^1, h_n)] \\ &\quad + [G(v_n^0, v_n^1, h_n) - G(u^0, u^1, h)], \end{aligned}$$

we can easily prove that $G_\varepsilon(u^0, u^1, h)$ converges to $G(u^0, u^1, h)$. The proof of Theorem 3.1 is now complete. \square

Using this result and Theorem 3.2, we can give the estimates on $c(\varepsilon)$.

THEOREM 3.3. *If Γ_0 satisfies (2.1), for $T > T_0$, there exist $c > 0$ and $\varepsilon_0 > 0$ depending only on the geometry of Ω and T , such that for all $(\phi^0, \phi^1) \in H_0^1(\Omega) \times L^2(\Omega)$, the solution ϕ of (2.3) satisfies*

$$\|\phi^0\|_{H_0^1(\Omega)}^2 + \|\phi^1\|_{L^2(\Omega)}^2 \leq c \left(\frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} \phi'^2(x, t) dx dt \right) \quad \forall \varepsilon \in]0, \varepsilon_0[.$$

Remark 3.2. This proves that $c(\varepsilon) = O(1/\varepsilon^3)$. This estimate is optimal. Indeed, consider the one-dimensional case and, for example, $\Omega =]0, 1[$. If $x_0 = 1$, we have $\Gamma_0 = \{0\}$ and $\omega_\varepsilon =]0, \varepsilon[$. The function $\phi(x, t) = \sin \pi t \sin \pi x$ is solution of (2.3), and $\|\nabla_\phi^0\|_{L^2(\Omega)}^2 + \|\phi^1\|_{L^2(\Omega)}^2 = \pi^2/2$, whereas

$$\int_0^T \int_0^\varepsilon \phi'^2(x, t) \, dx \, dt = \frac{1}{2} \left(\frac{T}{2} - \frac{\sin 2\pi T}{4\pi} \right) \left(\varepsilon - \frac{\sin 2\pi \varepsilon}{2\pi} \right) = O(\varepsilon^3).$$

Proof of Theorem 3.3. We argue by contradiction: Suppose Theorem 3.3 false. There exists then a sequence of nonnegative numbers $(\varepsilon_n)_n$ converging to 0 and sequences $(\tilde{\phi}_n^0)_n$ and $(\tilde{\phi}_n^1)_n$ of elements of $H_0^1(\Omega)$ and $L^2(\Omega)$ such that

$$(3.23) \quad \|\tilde{\phi}_n^0\|_{H_0^1(\Omega)}^2 + \|\tilde{\phi}_n^1\|_{L^2(\Omega)}^2 > n \left(\frac{1}{\varepsilon_n^3} \int_0^T \int_{\omega_{\varepsilon_n}} \tilde{\phi}_n'^2(x, t) \, dx \, dt \right),$$

where $\tilde{\phi}_n$ is the solution of (2.3) associated to the data $\tilde{\phi}_n^0$ and $\tilde{\phi}_n^1$.

We remark that $\tilde{\phi}_n$ cannot identically be equal to 0 because it would contradict (3.23). We then define

$$\phi_n^0 = \tilde{\phi}_n^0 / \|\tilde{\phi}_n^0\|_{H_0^1(\Omega)} \quad \text{and} \quad \phi_n^1 = \tilde{\phi}_n^1 / \|\tilde{\phi}_n^1\|_{L^2(\Omega)},$$

and we denote by ϕ_n the solution of (2.3) satisfying $\phi_n(0) = \phi_n^0$ and $\phi_n'(0) = \phi_n^1$. From (3.23), we have

$$(3.24) \quad \|\phi_n^0\|_{H_0^1(\Omega)}^2 + \|\phi_n^1\|_{L^2(\Omega)}^2 = 1$$

and

$$(3.25) \quad \frac{1}{\varepsilon_n^3} \int_0^T \int_{\omega_{\varepsilon_n}} \phi_n'^2(x, t) \, dx \, dt < \frac{1}{n}.$$

From (3.24), there exist $\phi^0 \in H_0^1(\Omega)$ and $\phi^1 \in L^2(\Omega)$ such that (after extraction of subsequences)

$$(3.26) \quad \phi_n^0 \text{ (resp., } \phi_n^1) \text{ weakly converges in } H_0^1(\Omega) \text{ (resp., } L^2(\Omega)) \text{ to } \phi^0 \text{ (resp., } \phi^1).$$

Thus, by (3.26), we get

$$(3.27) \quad \begin{aligned} \phi_n &\rightarrow \phi \text{ in } L^\infty(0, T; H_0^1(\Omega)) \text{ weak-}^*, \\ \phi_n' &\rightarrow \phi' \text{ in } L^\infty(0, T; L^2(\Omega)) \text{ weak-}^*, \end{aligned}$$

where ϕ is the solution of (2.3) associated with the data ϕ^0 and ϕ^1 .

LEMMA 3.4. *The limit ϕ is identically equal to 0.*

Proof of Lemma 3.4. The functions ϕ_n' satisfy the hypotheses of Theorem 3.1; hence from (3.21), we have $\partial\phi'/\partial\nu \in L^2(\Sigma_0)$, and, by (3.13) and (3.25), we have

$$\frac{\partial\phi'}{\partial\nu} = 0.$$

Using Lions's result recalled in Remark 3.1, we then have $\phi'(0) = \phi^1 = 0$ and $\phi''(0) = \Delta\phi^0 = 0$. As $\phi^0 \in H_0^1(\Omega)$, this implies $\phi^0 = \phi^1 = 0$; therefore $\phi = 0$.

To get a contradiction, we are first going to prove that (after extraction of a subsequence)

$$\frac{1}{\varepsilon_n^3} \int_0^T \int_{\omega_{\varepsilon_n}} \phi_n^2(x, t) \, dt \xrightarrow{n \rightarrow \infty} 0.$$

The proof of this result asks for several steps, and we begin with Lemma 3.5.

LEMMA 3.5. *There exists $c > 0$ depending only on Ω and T such that*

$$\frac{1}{\varepsilon_n^3} \int_{\omega_{\varepsilon_n}} \phi_n^2(x, t) dx \leq c \quad \forall n, \forall t \in [0, T].$$

Proof of Lemma 3.5. From (3.24) and Theorem 3.2 applied to ϕ_n and ε_n , there exists $c > 0$ depending only on Ω and T such that

$$(3.28) \quad \frac{1}{\varepsilon_n^3} \int_0^T \int_{\omega_{\varepsilon_n}} \phi_n^2(x, \tau) dx d\tau \leq c.$$

We then write

$$\phi_n(x, \tau) = \phi_n(x, t) + \int_t^\tau \phi_n'(x, s) ds;$$

thus

$$(3.29) \quad \phi_n^2(x, \tau) = \phi_n^2(x, t) + 2\phi_n(x, t) \int_t^\tau \phi_n'(x, s) ds + \left(\int_t^\tau \phi_n'(x, s) ds \right)^2.$$

Since $(\int_t^\tau \phi_n'(x, s) ds)^2 \geq 0$, putting (3.29) in (3.28), we obtain

$$(3.30) \quad \frac{T}{\varepsilon_n^3} \int_{\omega_{\varepsilon_n}} \phi_n^2(x, t) dx \leq c + \frac{2}{\varepsilon_n^3} \left| \int_0^T \int_{\omega_{\varepsilon_n}} \phi_n(x, t) \left(\int_t^\tau \phi_n'(x, s) ds \right) dx d\tau \right|.$$

On the other hand, for any $\gamma > 0$, we have

$$\begin{aligned} & \frac{2}{\varepsilon_n^3} \left| \int_0^T \int_{\omega_{\varepsilon_n}} \phi_n(x, t) \left(\int_t^\tau \phi_n'(x, s) ds \right) dx d\tau \right| \\ & \leq \frac{\gamma}{\varepsilon_n^3} \int_{\omega_{\varepsilon_n}} \phi_n^2(x, t) dx + \frac{1}{\gamma \varepsilon_n^3} \int_{\omega_{\varepsilon_n}} \left(\int_0^T \left(\int_t^\tau \phi_n'(x, s) ds \right) d\tau \right)^2 dx \\ & \leq \frac{\gamma}{\varepsilon_n^3} \int_{\omega_{\varepsilon_n}} \phi_n^2(x, t) dx + \frac{T^3}{\gamma \varepsilon_n^3} \int_{\omega_{\varepsilon_n}} \int_0^T \phi_n'^2(x, s) dx ds. \end{aligned}$$

Reporting in (3.30), we get

$$\begin{aligned} \frac{T}{\varepsilon_n^3} (1 - \gamma) \int_{\omega_{\varepsilon_n}} \phi_n^2(x, t) dx & \leq c + \frac{T^3}{\gamma \varepsilon_n^3} \int_{\omega_{\varepsilon_n}} \int_0^T \phi_n'^2(x, s) dx ds \\ & \leq c + \frac{T^3}{n\gamma}, \text{ from (3.25).} \end{aligned}$$

We choose $\gamma < 1$ to obtain Lemma 3.5.

We then consider

$$\psi_n(x, t) = \int_0^t \phi_n(x, s) ds + s_n$$

with $\Delta S_n = \phi_n^1$, $s_n \in H^2 \cap H_0^1(\Omega)$.

The function ψ_n is solution of (2.3) with initial data $\psi_n^0 = s_n$ in $H^2 \cap H_0^1(\Omega)$ and $\psi_n^1 = \phi_n^0$ in $H_0^1(\Omega)$.

If we take into account (3.24) and Lemma 3.4, we have

$$\phi_n^0 \rightarrow 0 \quad \text{in } H_0^1(\Omega) \text{ weak,}$$

$$\phi_n^1 \rightarrow 0 \quad \text{in } L^2(\Omega) \text{ weak;}$$

hence, by compact injection of $H_0^1(\Omega)$ in $L^2(\Omega)$ and of $L^2(\Omega)$ in $H^{-1}(\Omega)$,

$$(3.31) \quad \begin{aligned} \psi'_n(0) &= \phi_n^0 \rightarrow 0 \quad \text{in } L^2(\Omega) \text{ strong,} \\ \psi_n(0) &= s_n \rightarrow 0 \quad \text{in } H_0^1(\Omega) \text{ strong;} \end{aligned}$$

so ψ_n strongly converges to 0 in $C([0, T]; H_0^1(\Omega))$, and ψ'_n strongly converges to 0 in $C([0, T]; L^2(\Omega))$.

LEMMA 3.6. *There exists a nonnegative real number I and a subsequence $(n_k)_k$ such that*

$$\frac{1}{\varepsilon_{n_k}^3} \int_{\omega_{\varepsilon_{n_k}}} \psi_{n_k}^2(x, t) dx \xrightarrow[k \rightarrow \infty]{} I \quad \forall t \in [0, T]$$

(the subsequence is the same for all t).

Proof of Lemma 3.6. Using the same argument as in Lemma 3.5 and taking into account (3.31), we can prove that there exists $c > 0$ independent of n such that

$$(3.32) \quad \frac{1}{\varepsilon_n^3} \int_{\omega_{\varepsilon_n}} \psi_n^2(x, t) dx \leq c \quad \forall n, \forall t \in [0, T].$$

Using (3.32) for $t = 0$, we obtain the existence of a subsequence $(n_k)_k$ and a number I with $I \geq 0$ such that

$$(3.33) \quad \frac{1}{\varepsilon_{n_k}^3} \int_{\omega_{\varepsilon_{n_k}}} \psi_{n_k}^2(x, 0) dx \xrightarrow[k \rightarrow \infty]{} I.$$

By an integration by parts in time, we get

$$\begin{aligned} & \frac{1}{\varepsilon_{n_k}^3} \int_{\omega_{\varepsilon_{n_k}}} \psi_{n_k}^2(x, t) dx - \frac{1}{\varepsilon_{n_k}^3} \int_{\omega_{\varepsilon_{n_k}}} \psi_{n_k}^2(x, 0) dx \\ &= \frac{2}{\varepsilon_{n_k}^3} \int_0^t \int_{\omega_{\varepsilon_{n_k}}} \psi_{n_k}(x, s) \phi_{n_k}(x, s) dx ds, \end{aligned}$$

and, using Hölder's inequality,

$$\begin{aligned} \left| \frac{1}{\varepsilon_{n_k}^3} \int_0^t \int_{\omega_{\varepsilon_{n_k}}} \psi_{n_k}(x, s) \phi_{n_k}(x, s) dx ds \right| &\leq \left(\frac{1}{\varepsilon_{n_k}^3} \int_0^t \int_{\omega_{\varepsilon_{n_k}}} \psi_{n_k}^2(x, s) dx ds \right)^{1/2} \\ &\quad \cdot \left(\frac{1}{\varepsilon_{n_k}^3} \int_0^t \int_{\omega_{\varepsilon_{n_k}}} \phi_{n_k}^2(x, s) dx ds \right)^{1/2}. \end{aligned}$$

From Theorem 3.2, (3.24), and (3.31), for any $t > 0$, we have

$$\frac{1}{\varepsilon_{n_k}^3} \int_0^t \int_{\omega_{\varepsilon_{n_k}}} \phi_{n_k}^2(x, s) dx ds \leq c$$

and

$$(3.34) \quad \frac{1}{\varepsilon_{n_k}^3} \int_0^t \int_{\omega_{\varepsilon_{n_k}}} \psi_{n_k}^2(x, s) dx ds \leq c(\|\psi_{n_k}^0\|_{H_0^1(\Omega)}^2 + \|\psi_{n_k}^1\|_{L^2(\Omega)}^2) \xrightarrow[k \rightarrow \infty]{} 0.$$

We then deduce that

$$\frac{1}{\varepsilon_{n_k}^3} \int_0^t \int_{\omega_{\varepsilon_{n_k}}} \psi_{n_k}(x, s) \phi_{n_k}(x, s) dx ds \xrightarrow[k \rightarrow \infty]{} 0 \quad \forall t \in [0, T],$$

which proves

$$\frac{1}{\varepsilon_{n_k}^3} \int_{\omega_{\varepsilon_{n_k}}} \psi_{n_k}^2(x, t) dx \xrightarrow{k \rightarrow \infty} I, \quad \forall t \in [0, T].$$

LEMMA 3.7. We have $I = 0$.

Proof of Lemma 3.7. We write

$$f_k(t) = \frac{1}{\varepsilon_{n_k}^3} \int_{\omega_{\varepsilon_{n_k}}} \psi_{n_k}^2(x, t) dx.$$

From Lemma 3.6, we have

$$f_k \xrightarrow{k \rightarrow \infty} I \quad \text{almost everywhere on } [0, T],$$

and, by (3.32),

$$|f_k(t)| \leq c, \quad \text{where } c \text{ does not depend on } k \text{ and } t.$$

From Lebesgue's theorem, we then deduce that

$$\int_0^T f_k(t) dt \xrightarrow{k \rightarrow \infty} IT.$$

However, on the other hand, by (3.34), we have

$$\int_0^T f_k(t) dt = \frac{1}{\varepsilon_{n_k}^3} \int_0^T \int_{\omega_{\varepsilon_{n_k}}} \psi_{n_k}^2(x, t) dx dt \xrightarrow{k \rightarrow \infty} 0,$$

which proves that $I = 0$. \square

We now can prove the following lemma.

LEMMA 3.8. We have

$$\frac{1}{\varepsilon_{n_k}^3} \int_0^T \int_{\omega_{\varepsilon_{n_k}}} \phi_{n_k}^2(x, t) dx dt \xrightarrow{k \rightarrow \infty} 0.$$

Proof of Lemma 3.8. From Hölder's inequality, we get

$$\begin{aligned} \left| \frac{1}{\varepsilon_{n_k}^3} \int_0^T \int_{\omega_{\varepsilon_{n_k}}} \psi_{n_k}(x, t) \phi'_{n_k}(x, t) dx dt \right|^2 &\leq \left(\frac{1}{\varepsilon_{n_k}^3} \int_0^T \int_{\omega_{\varepsilon_{n_k}}} \phi'^2_{n_k}(x, t) dx dt \right) \\ &\quad \cdot \left(\frac{1}{\varepsilon_{n_k}^3} \int_0^T \int_{\omega_{\varepsilon_{n_k}}} \psi_{n_k}^2(x, t) dx dt \right). \end{aligned}$$

Each term in the right-hand side tends to 0 when k goes to ∞ ; thus

$$\left| \frac{1}{\varepsilon_{n_k}^3} \int_0^T \int_{\omega_{\varepsilon_{n_k}}} \psi_{n_k}(x, t) \phi'_{n_k}(x, t) dx dt \right|^2 \xrightarrow{k \rightarrow \infty} 0.$$

However, by an integration by parts in time, as $\phi_{n_k} = \psi'_{n_k}$

$$\begin{aligned} \frac{1}{\varepsilon_{n_k}^3} \int_0^T \int_{\omega_{\varepsilon_{n_k}}} \psi_{n_k}(x, t) \phi'_{n_k}(x, t) dx dt &= \frac{1}{\varepsilon_{n_k}^3} \int_{\omega_{\varepsilon_{n_k}}} \phi_{n_k}(x, T) \psi_{n_k}(x, T) dx \\ &\quad - \frac{1}{\varepsilon_{n_k}^3} \int_{\omega_{\varepsilon_{n_k}}} \phi_{n_k}(x, 0) \psi_{n_k}(x, 0) dx \\ &\quad - \frac{1}{\varepsilon_{n_k}^3} \int_0^T \int_{\omega_{\varepsilon_{n_k}}} \phi_{n_k}^2(x, t) dx dt. \end{aligned}$$

From Lemmas 3.5–3.7, again using Hölder's inequality, we can show that both of the boundary terms in time go to 0 when k goes to infinity. We then deduce Lemma 3.8.

Now, using a Zuazua result (see [7, Lemma 2.4, p. 413]), we have

$$\begin{aligned} \|\phi_{n_k}^0\|_{H_0^1(\Omega)}^2 + \|\phi_{n_k}^1\|_{L^2(\Omega)}^2 &\leq c \left(\frac{1}{\varepsilon_{n_k}^3} \int_0^T \int_{\omega_{\varepsilon_{n_k}}} \phi_{n_k}^2(x, t) \, dx \, dt \right. \\ &\quad \left. + \frac{1}{\varepsilon_{n_k}^3} \int_0^T \int_{\omega_{\varepsilon_{n_k}}} \phi_{n_k}'^2(x, t) \, dx \, dt \right). \end{aligned}$$

By (3.25) and Lemma 3.8, we then get

$$\|\phi_{n_k}^0\|_{H_0^1(\Omega)}^2 + \|\phi_{n_k}^1\|_{L^2(\Omega)}^2 \xrightarrow{k \rightarrow \infty} 0,$$

which contradicts (3.24) and finishes the proof of Theorem 3.3.

We can easily prove that Theorem 3.3 is equivalent to Theorem 3.4.

THEOREM 3.4. *If Γ_0 satisfies (2.1), for $T > T_0$, there exists $d > 0$, such that for all $(\varphi^0, \varphi^1) \in L^2(\Omega) \times H^{-1}(\Omega)$, the solution φ of (2.3) satisfies*

$$\|\varphi^0\|_{L^2(\Omega)}^2 + \|\varphi^1\|_{H^{-1}(\Omega)}^2 \leq d \left(\frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} \varphi^2(x, t) \, dx \, dt \right).$$

This last theorem is essential to obtain the estimates on the controls and study the passage to the limit when ε goes to 0 in problem (2.2), which we now recall.

4. Passing to the limit in the exact controllability problems. We saw in § 2 that if Γ_0 satisfies (2.1), for $T > T_0$, and for any couple of initial data $(y^0, y^1) \in H_0^1(\Omega) \times L^2(\Omega)$, the solution ψ_ε of the exact controllability problem given by HUM, is defined by

$$\begin{aligned} \psi_\varepsilon'' - \Delta \psi_\varepsilon &= \tilde{\varphi}_\varepsilon \chi_{\omega_\varepsilon \times (0, T)} \quad \text{in } Q, \\ \psi_\varepsilon &= 0 \quad \text{on } \Sigma, \\ \psi_\varepsilon(0) &= y^0 \quad \text{and} \quad \psi_\varepsilon'(0) = y^1, \\ \psi_\varepsilon(T) &= 0 \quad \text{and} \quad \psi_\varepsilon'(T) = 0, \end{aligned}$$

where $\tilde{\varphi}_\varepsilon$ is the solution of (2.3) with initial data $(\tilde{\varphi}_\varepsilon^0, \tilde{\varphi}_\varepsilon^1) \in L^2(\Omega) \times H^{-1}(\Omega)$ satisfying

$$(4.1) \quad -(\tilde{\varphi}_\varepsilon^0, y^1) + \langle \tilde{\varphi}_\varepsilon^1, y^0 \rangle = \int_0^T \int_{\omega_\varepsilon} \tilde{\varphi}_\varepsilon^2(x, t) \, dx \, dt.$$

From Theorem 3.4, we deduce the following estimates.

THEOREM 4.1. $\tilde{\varphi}_\varepsilon^0, \tilde{\varphi}_\varepsilon^1$, and $\tilde{\varphi}_\varepsilon$ satisfy

$$\|(\tilde{\varphi}_\varepsilon^0, \tilde{\varphi}_\varepsilon^1)\|_{L^2 \times H^{-1}} = O\left(\frac{1}{\varepsilon^3}\right) \quad \text{and} \quad \int_0^T \int_{\omega_\varepsilon} \tilde{\varphi}_\varepsilon^2(x, t) \, dx \, dt = O\left(\frac{1}{\varepsilon^3}\right).$$

Proof of Theorem 4.1. By definition of Λ_ε , (4.1) implies

$$(4.2) \quad \int_0^T \int_{\omega_\varepsilon} \tilde{\varphi}_\varepsilon^2(x, t) \, dx \, dt \leq \|(\tilde{\varphi}_\varepsilon^0, \tilde{\varphi}_\varepsilon^1)\|_{L^2 \times H^{-1}} \|(y^0, y^1)\|_{H_0^1 \times L^2}.$$

On the other hand, by Theorem 3.4, we have

$$(4.3) \quad \|(\tilde{\varphi}_\varepsilon^0, \tilde{\varphi}_\varepsilon^1)\|_{L^2 \times H^{-1}}^2 \leq c \left(\frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} \tilde{\varphi}_\varepsilon^2(x, t) \, dx \, dt \right),$$

where c does not depend on ε . Taking into account (4.2) and (4.3), we obtain

$$\begin{aligned} \|(\tilde{\varphi}_\varepsilon^0, \tilde{\varphi}_\varepsilon^1)\|_{L^2 \times H^1}^2 &\leq c \left(\frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} \tilde{\phi}_\varepsilon^2(x, t) \, dx \, dt \right) \\ &\leq \frac{c}{\varepsilon^3} \|(\tilde{\phi}_\varepsilon^0, \tilde{\phi}_\varepsilon^1)\|_{L^2 \times H^{-1}} \|(y^0, y^1)\|_{H_0^1 \times L^2}, \end{aligned}$$

and this gives the result of Theorem 4.1.

We then introduce $\varphi_\varepsilon^0 = \varepsilon^3 \tilde{\phi}_\varepsilon^0$ and $\varphi_\varepsilon^1 = \varepsilon^3 \tilde{\phi}_\varepsilon^1$. The function $\varphi_\varepsilon = \varepsilon^3 \tilde{\phi}_\varepsilon$ is solution of (2.3) associated with the initial data φ_ε^0 and φ_ε^1 , and we have from Theorem 4.1

$$(4.4) \quad \|(\varphi_\varepsilon^0, \varphi_\varepsilon^1)\|_{L^2 \times H^{-1}} = O(1), \quad \frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} \varphi_\varepsilon^2(x, t) \, dx \, dt = O(1).$$

Thus, there exists $\varphi^0 \in L^2(\Omega)$ and $\varphi^1 \in H^{-1}(\Omega)$ such that (after extraction of subsequences)

$$\varphi_\varepsilon^0 \rightarrow \varphi^0 \text{ in } L^2(\Omega) \text{ weak and } \varphi_\varepsilon^1 \rightarrow \varphi^1 \text{ in } H^{-1}(\Omega) \text{ weak.}$$

Define φ as the solution of (2.3) with $\varphi(0) = \varphi^0$ and $\varphi'(0) = \varphi^1$. We know then that $\varphi_\varepsilon \rightarrow \varphi$ in $L^\infty(0, T; L^2(\Omega))$ weak-*, and, by Theorem 3.1, $\partial\varphi/\partial\nu \in L^2(\Sigma_0)$. As Γ_0 satisfies (2.1), from Lions's result, $\varphi^0 \in H_0^1(\Omega)$, $\varphi^1 \in L^2(\Omega)$, and φ is a solution of (2.3) with finite energy. This regularity result that we obtain for the limit φ is essential to obtain the same spaces of initial data for the exact boundary controllability problem that we will obtain as limit of problem (4.5) just above.

With these new notations, ψ_ε satisfies

$$\begin{aligned} (4.5) \quad \psi_\varepsilon'' - \Delta \psi_\varepsilon &= \frac{1}{\varepsilon^3} \varphi_\varepsilon \chi_{\omega_\varepsilon \times (0, T)} \quad \text{in } Q, \\ \psi_\varepsilon &= 0 \quad \text{on } \Sigma, \\ \psi_\varepsilon(0) &= y^0, \psi_\varepsilon'(0) = y^1 \quad \text{and} \quad \psi_\varepsilon(T) = 0, \psi_\varepsilon'(T) = 0. \end{aligned}$$

Remark 4.1. We are led to study the convergence of solutions of wave equations with singular right-hand sides and homogeneous boundary conditions. This problem can be stated independently of the exact controllability context and for other equations associated to different operators (for example, the Schrödinger or the dynamic plate equation).

The estimates given by Theorem 4.1 and (4.4) permit us to prove the following theorem.

THEOREM 4.2. *If Γ_0 satisfies (2.1), for $T > T_0$ and for every $(y^0, y^1) \in H_0^1(\Omega) \times L^2(\Omega)$, the solutions ψ_ε of (4.5) converge (after extraction of a subsequence) for the weak-* topology of $L^\infty(0, T; L^2(\Omega))$ to the solution ψ of the following exact controllability problem:*

$$\begin{aligned} (4.6) \quad \psi'' - \Delta \psi &= 0 \quad \text{in } Q, \\ \psi &= -\frac{1}{3} \frac{\partial \varphi}{\partial \nu} \in L^2(\Sigma_0) \text{ on } \Sigma_0 \quad \text{and} \quad \psi = 0 \text{ on } \Sigma - \Sigma_0, \\ \psi(0) &= y^0 \quad \text{and} \quad \psi'(0) = y^1, \\ \psi(T) &= 0 \quad \text{and} \quad \psi'(T) = 0, \end{aligned}$$

where φ is the limit (in a weak sense) of φ_ε .

Remark 4.2. Since the boundary condition is not conserved, we cannot hope for a convergence in $L^\infty(0, T; H_0^1(\Omega))$. This explains why we are going to look at ψ_ε as a solution defined by transposition.

Proof of Theorem 4.2. For $(u^0, u^1, h) \in H_0^1(\Omega) \times L^2(\Omega) \times L^1(0, T; L^2(\Omega))$, let u be the solution of (3.2) associated to these data. We multiply (4.5) by u , and we integrate by parts to obtain

$$(4.7) \quad ((\psi_\varepsilon, h)) = \frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} \varphi_\varepsilon(x, t) u(x, t) \, dx \, dt + \langle y^1, u^0 \rangle - \langle y^0, u^1 \rangle,$$

where $((\cdot, \cdot))$ denotes the duality $L^\infty(0, T; L^2(\Omega)), L^1(0, T; L^2(\Omega))$.

We introduce the linear forms

$$G_\varepsilon : H_0^1(\Omega) \times L^2(\Omega) \times L^1(0, T; L^2(\Omega)) \rightarrow \mathbb{R},$$

$$(u^0, u^1, h) \rightarrow \frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} \varphi_\varepsilon(x, t) u(x, t) \, dx \, dt.$$

From (4.4) and Theorem 3.1, $(G_\varepsilon)_\varepsilon$ is bounded in $H^{-1}(\Omega) \times L^2(\Omega) \times L^\infty(0, T; L^2(\Omega))$ and, after extraction of a subsequence, $(G_\varepsilon)_\varepsilon$ converges for the weak-* topology of this space to

$$G(u^0, h^1, h) = \frac{1}{3} \int_{\Sigma_0} \frac{\partial \varphi}{\partial \nu}(y, t) \frac{\partial u}{\partial \nu}(y, t) \, dy \, dt.$$

Equation (4.7) then proves that $(\psi_\varepsilon)_\varepsilon$ is bounded in $L^\infty(0, T; L^2(\Omega))$ and therefore (after extraction of a subsequence) converge in $L^\infty(0, T; L^2(\Omega))$ weak-* to an element ψ of $L^\infty(0, T; L^2(\Omega))$. Passing to the limit in (4.7), we then deduce that ψ is solution of

$$(4.8) \quad ((\psi, h)) = \frac{1}{3} \int_{\Sigma_0} \frac{\partial \varphi}{\partial \nu}(y, t) \frac{\partial u}{\partial \nu}(y, t) \, dy \, dt + \langle y^1, u^0 \rangle - \langle y^0, u^1 \rangle.$$

The interpretation of (4.8), for which we can refer to [7], shows that ψ is the solution of (4.6). Indeed, (4.8) gives the exact meaning of (4.6) using the transposition method (see [7]).

Remark 4.3. By definition of φ_ε , we have

$$(4.9) \quad -\langle y^1, \varphi_\varepsilon^0 \rangle + \langle \varphi_\varepsilon^1, y^0 \rangle = \frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} \varphi_\varepsilon^2(x, t) \, dx \, dt.$$

Thus (after extraction of a subsequence)

$$\frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} \varphi_\varepsilon^2(x, t) \, dx \, dt \xrightarrow{\varepsilon \rightarrow 0} -\langle y^1, \varphi^0 \rangle + \langle \varphi^1, y^0 \rangle = -\langle y^1, \varphi^0 \rangle + (\varphi^1, y^0).$$

On the other hand,

$$(4.10) \quad -\langle y^1, \varphi^0 \rangle + (\varphi^1, y^0) = \frac{1}{3} \int_{\Sigma_0} \left| \frac{\partial \varphi}{\partial \nu}(y, t) \right|^2 \, dy \, dt.$$

Hence, in this case

$$\lim_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} \varphi_\varepsilon^2(x, t) \, dx \, dt = \frac{1}{3} \int_{\Sigma_0} \left| \frac{\partial \varphi}{\partial \nu}(y, t) \right|^2 \, dy \, dt.$$

Furthermore, using Lions's results concerning the exact boundary controllability of the wave equation, for which we can refer to [6] or [7], (4.10) shows the uniqueness

of the data (φ^0, φ^1) ; thus all the sequence $(\varphi_\varepsilon^0, \varphi_\varepsilon^1)$ to (φ^0, φ^1) and, by the way, all the sequence ψ_ε converges to the solutions of (4.6).

Remark 4.4. The exact controllability of the wave equation when the control acts on Σ_0 has been studied by Lions in [1] or [2], and he proved that the control given by HUM belongs to $L^2(\Sigma_0)$ for every initial data $(y^0, y^1) \in L^2(\Omega) \times H^{-1}(\Omega)$. We still must show that we can also reach these spaces of initial data. So, let us take $(y^0, y^1) \in L^2(\Omega) \times H^{-1}(\Omega)$. We consider sequences $y_n^0 \in H_0^1(\Omega)$ and $y_n^1 \in L^2(\Omega)$ such that $y_n^0 \rightarrow y^0$ in $L^2(\Omega)$ and $y_n^1 \rightarrow y^1$ in $H^{-1}(\Omega)$. For every n , let φ_n and ψ_n be the solution of Theorem 4.2. If we apply (4.10), we get

$$-(y_n^1, \varphi_n^0) + (\varphi_n^1, y_n^0) = \frac{1}{3} \int_{\Sigma_0} \left| \frac{\partial \varphi_n}{\partial \nu}(y, t) \right|^2 dy dt.$$

Hence, using Lions's result, we obtain

$$\|(\varphi_n^0, \varphi_n^1)\|_{H_0^1 \times L^2}^2 \leq \frac{c}{3} \int_{\Sigma_0} \left| \frac{\partial \varphi_n}{\partial \nu}(y, t) \right|^2 dy dt \leq c \|(y_n^0, y_n^1)\|_{L^2 \times H^{-1}} \|(\varphi_n^0, \varphi_n^1)\|_{H_0^1 \times L^2}.$$

We then deduce that there exists $\varphi^0 \in H_0^1(\Omega)$ and $\varphi^1 \in L^2(\Omega)$ such that (after extraction of subsequences) φ_n^0 weakly converges to φ^0 in $H_0^1(\Omega)$, φ_n^1 weakly converges to φ^1 in $L^2(\Omega)$, and

$$\frac{\partial \varphi_n}{\partial \nu} \rightarrow \frac{\partial \varphi}{\partial \nu} \quad \text{in } L^2(\Sigma_0) \text{ weak,}$$

where φ is the solution of (2.3) associated to the initial data φ^0 and φ^1 .

To conclude, we use the following result of Lions (see [7]). For $(z^0, z^1, v) \in L^2(\Omega) \times H^{-1} \times L^2(\Sigma)$, there exists one and only one solution z in $C(0, T; L^2(\Omega)) \cap C^1(0, T; H^{-1}(\Omega))$ of

$$\begin{aligned} z'' - \Delta z &= 0 \quad \text{in } Q, \\ z &= v \quad \text{on } \Sigma, \\ z(0) &= z^0 \quad \text{and} \quad z'(0) = z^1. \end{aligned}$$

Furthermore, the mapping

$$(z^0, z^1, v) \in L^2(\Omega) \times H^{-1} \times L^2(\Sigma) \rightarrow z \in C(0, T; L^2(\Omega)) \cap C^1(0, T; H^{-1}(\Omega))$$

is continuous.

We can easily prove that, passing to the limit when n goes to infinity, we obtain the solution of (4.6), where the initial data (y^0, y^1) belong to $L^2(\Omega) \times H^{-1}(\Omega)$.

5. Changing the space of initial data. We consider here the case of the same internal exact controllability problem when the initial data (y^0, y^1) belong to $L^2(\Omega) \times H^{-1}(\Omega)$. Zuazua proved in this case that if Γ_0 satisfies (2.1), for $T > T_0$, the solution given by HUM of the exact controllability problem when the control acts on Q_ε is defined by the following system:

$$\begin{aligned} \psi_\varepsilon'' - \Delta \psi_\varepsilon &= -\frac{\partial}{\partial t}(\tilde{\varphi}'_\varepsilon) \chi_{\omega_\varepsilon \times (0, T)} \quad \text{in } Q, \\ \psi_\varepsilon &= 0 \quad \text{on } \Sigma, \\ \psi_\varepsilon(0) &= y^0 \quad \text{and} \quad \psi'_\varepsilon(0) = y^1, \\ \psi_\varepsilon(T) &= 0 \quad \text{and} \quad \psi'_\varepsilon(T) = 0, \end{aligned}$$

where $\tilde{\varphi}_\varepsilon$ is the solution of (2.3) whose initial data $\tilde{\varphi}_\varepsilon^0$ and $\tilde{\varphi}_\varepsilon^1$ satisfy

$$(5.1) \quad -\langle y^1, \tilde{\varphi}_\varepsilon^0 \rangle + \langle y^0, \tilde{\varphi}_\varepsilon^1 \rangle = \int_0^T \int_{\omega_\varepsilon} \tilde{\varphi}_\varepsilon'^2(x, t) \, dx \, dt,$$

and where $-\partial/\partial t(\tilde{\varphi}_\varepsilon')\chi_{\omega_\varepsilon \times (0, T)}$ belongs to $(H^1(0, T; L^2(\omega_\varepsilon)))'$ and is defined by

$$\left\langle -\frac{\partial}{\partial t}(\tilde{\varphi}_\varepsilon')\chi_{\omega_\varepsilon \times (0, T)}, u \right\rangle = \int_{Q_\varepsilon} \varphi'(x, t) u'(x, t) \, dx \, dt, \quad \text{for } u \in H^1(0, T; L^2(\omega_\varepsilon)).$$

Using Theorem 3.3, it follows that

$$\begin{aligned} \|(\tilde{\varphi}_\varepsilon^0, \tilde{\varphi}_\varepsilon^1)\|_{H_0^1 \times L^2}^2 &\leq c \left(\frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} \tilde{\varphi}_\varepsilon'^2(x, t) \, dx \, dt \right) \\ &\leq \frac{c}{\varepsilon^3} \|(\tilde{\varphi}_\varepsilon^0, \tilde{\varphi}_\varepsilon^1)\|_{H_0^1 \times L^2} \|(y^0, y^1)\|_{L^2 \times H^{-1}}. \end{aligned}$$

This gives the estimates

$$(5.2) \quad \|(\tilde{\varphi}_\varepsilon^0, \tilde{\varphi}_\varepsilon^1)\|_{H_0^1 \times L^2} = O\left(\frac{1}{\varepsilon^3}\right) \quad \text{and} \quad \int_0^T \int_{\omega_\varepsilon} \tilde{\varphi}_\varepsilon'^2(x, t) \, dx \, dt = O\left(\frac{1}{\varepsilon^3}\right).$$

We then write $\varphi_\varepsilon^0 = \varepsilon^3 \tilde{\varphi}_\varepsilon^0$ and $\varphi_\varepsilon^1 = \varepsilon^3 \tilde{\varphi}_\varepsilon^1$. The function $\varphi_3 = \varepsilon^3 \tilde{\varphi}_\varepsilon$ is solution of (2.3), and its initial data φ_ε^0 and φ_ε^1 satisfy by (5.2)

$$(5.3) \quad \begin{aligned} \|(\varphi_\varepsilon^0, \varphi_\varepsilon^1)\|_{H_0^1 \times L^2} &= O(1), \\ \frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} \varphi_\varepsilon'^2(x, t) \, dx \, dt &= O(1). \end{aligned}$$

There exists $\varphi^0 \in H_0^1(\Omega)$ and $\varphi^1 \in L^2(\Omega)$ such that (after extraction of subsequences)

$$\varphi_\varepsilon^0 \rightarrow \varphi^0 \text{ in } H_0^1(\Omega) \text{ weak} \quad \text{and} \quad \varphi_\varepsilon^1 \rightarrow \varphi^1 \text{ in } L^2(\Omega) \text{ weak}.$$

Let φ be the solution of (2.3) with $\varphi(0) = \varphi^0$ and $\varphi'(0) = \varphi^1$. We then know that $\varphi_\varepsilon \rightarrow \varphi$ in $L^\infty(0, T; H_0^1(\Omega))$ weak-* and $\varphi'_\varepsilon \rightarrow \varphi'$ in $L^\infty(0, T; L^2(\Omega))$ weak-*. Applying Theorem 3.1 to φ'_ε , we get $\partial \varphi' / \partial \nu \in L^2(\Sigma_0)$. Since Γ_0 satisfies (2.1), we then deduce that $\varphi^1 \in H_0^1(\Omega)$ and $\varphi''(0) = \Delta \varphi^0 \in L^2(\Omega)$. This implies that $\varphi^0 \in H^2(\Omega) \cap H_0^1(\Omega)$ and that φ is a strong solution of (2.3).

With these new notations, ψ_ε is solution of

$$(5.3) \quad \begin{aligned} \psi_\varepsilon'' - \Delta \psi_\varepsilon &= -\frac{1}{\varepsilon^3} \frac{\partial}{\partial t}(\varphi'_\varepsilon) \chi_{\omega_\varepsilon \times (0, T)} \quad \text{in } Q, \\ \psi_\varepsilon &= 0 \quad \text{on } \Sigma, \\ \psi_\varepsilon(0) &= y^0 \quad \text{and} \quad \psi'_\varepsilon(0) = y^1, \\ \psi_\varepsilon(T) &= 0 \quad \text{and} \quad \psi'_\varepsilon(T) = 0, \end{aligned}$$

and we have the following theorem.

THEOREM 5.1. *If Γ_0 satisfies (2.1), for $T > T_0$ and for every couple of initial data $(y^0, y^1) \in L^2(\Omega) \times H^{-1}(\Omega)$, the solutions ψ_ε of (5.3) converge (after extraction of a subsequence) for the weak-* topology of $L^\infty(0, T; H^{-1}(\Omega))$ to the solution ψ of the*

following exact controllability problem:

$$\begin{aligned}
 (5.4) \quad & \psi'' - \Delta \psi = 0 \quad \text{in } Q, \\
 & \psi = \frac{1}{3} \frac{\partial}{\partial t} \left(\frac{\partial \varphi'}{\partial \nu} \right) \in (H^1(0, T; L^2(\Gamma_0)))' \text{ on } \Sigma_0 \quad \text{and} \quad \psi = 0 \text{ on } \Sigma - \Sigma_0, \\
 & \psi(0) = y^0 \quad \text{and} \quad \psi'(0) = y^1, \\
 & \psi(T) = 0 \quad \text{and} \quad \psi'(T) = 0,
 \end{aligned}$$

where φ is the limit (in a weak sense) of φ_ε .

Remark 5.1. $(\partial/\partial t)(\partial\varphi'/\partial\nu)$ denotes the element of $(H^1(0, T; L^2(\Gamma_0)))'$ defined by

$$\left[\frac{\partial}{\partial t} \left(\frac{\partial \varphi'}{\partial \nu} \right), v \right] = - \int_{\Sigma_0} \frac{\partial \varphi'}{\partial \nu}(y, t) \frac{\partial v'}{\partial \nu}(y, t) dy dt \quad \forall v \in H^1(0, T; L^2(\Gamma_0)),$$

where $[\cdot]$ denotes the duality $(H^1(0, T; L^2(\Gamma_0)))', H^1(0, T; L^2(\Gamma_0))$.

Proof of Theorem 5.1. We consider ψ_ε as a solution of the wave equation with initial data in $H^{-1}(\Omega) \times (H^2(\Omega) \cap H_0^1(\Omega))'$ (the properties of these solutions are described in [7]). For $(u^0, u^1, h) \in (H^2(\Omega) \cap H_0^1(\Omega)) \times H_0^1(\Omega) \times L^1(0, T; H_0^1(\Omega))$ let u be the solution of (3.2) associated with these data. We have

$$(5.5) \quad \langle \psi_\varepsilon, h \rangle = \frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} \varphi'_\varepsilon(x, t) u'(x, t) dx dt + \{y^0, u^1\} - \langle y^0, u^1 \rangle,$$

here $\langle \cdot \rangle$ (respectively, $\{ \cdot \}$) denotes the duality $L^\infty(0, T; H^{-1}(\Omega)), L^1(0, T; H_0^1(\Omega))$ (respectively, $(H^2(\Omega) \cap H_0^1(\Omega))', H^2(\Omega) \cap H_0^1(\Omega)$).

We introduce the linear forms defined by

$$\begin{aligned}
 G_\varepsilon : H^2(\Omega) \cap H_0^1(\Omega) \times H_0^1(\Omega) \times L^1(0, T; H_0^1(\Omega)) &\rightarrow \mathbb{R}, \\
 (u^0, u^1, h) &\rightarrow \frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} \varphi'_\varepsilon(x, t) u'(x, t) dx dt.
 \end{aligned}$$

The study of the convergence of these forms is similar to the case that we saw in Theorem 3.1. To prove that they are bounded in $(H^2(\Omega) \cap H_0^1(\Omega))' \times H^{-1}(\Omega) \times L^\infty(0, T; H^{-1}(\Omega))$, we use a result given in [5] concerning the behavior near the boundary for solutions of the wave equation with data in $H^2(\Omega) \cap H_0^1(\Omega) \times H_0^1(\Omega) \times L^1(0, T; H_0^1(\Omega))$. After extraction of a subsequence, $(G_\varepsilon)_\varepsilon$ converges in $(H^2(\Omega) \cap H_0^1(\Omega))' \times H^{-1}(\Omega) \times L^\infty(0, T; H^{-1}(\Omega))$ weak-* to an element G . To describe G , we first consider (u^0, u^1, h) in $(H^2(\Omega) \cap H_0^1(\Omega)) \times H_0^1(\Omega) \times D(0, T; H_0^1(\Omega))$. Then applying Theorem 3.1 to φ'_ε , we can prove that, for (u^0, u^1, h) in $(H^2(\Omega) \cap H_0^1(\Omega)) \times H_0^1(\Omega) \times D(0, T; H_0^1(\Omega))$, we have

$$G(u^0, u^1, h) = \frac{1}{3} \int_{\Sigma_0} \frac{\partial \varphi'}{\partial \nu}(y, t) \frac{\partial u'}{\partial \nu}(y, t) dy dt.$$

By a density argument, we then deduce that, for all $(u^0, u^1, h) \in (H^2(\Omega) \cap H_0^1(\Omega)) \times H_0^1(\Omega) \times L^1(0, T; H_0^1(\Omega))$,

$$G(u^0, u^1, h) = \frac{1}{3} \int_{\Sigma_0} \frac{\partial \varphi'}{\partial \nu}(y, t) \frac{\partial u'}{\partial \nu}(y, t) dy dt.$$

Having this result, (5.5) proves $(\psi_\varepsilon)_\varepsilon$ is bounded in $L^\infty(0, T; H^{-1}(\Omega))$, and therefore (after extraction of a subsequence) it converges in $L^\infty(0, T; H^{-1}(\Omega))$ weak-* to an

element ψ of $L^\infty(0, T; H^{-1}(\Omega))$. Passing to the limit in (5.5), we then obtain that ψ is solution of

$$(5.6) \quad \langle\langle \psi, h \rangle\rangle = \frac{1}{3} \int_{\Sigma_0} \frac{\partial \varphi'}{\partial \nu}(y, t) \frac{\partial u'}{\partial \nu}(y, t) dy dt + \{y^0, u^1\} - \langle y^0, u^1 \rangle,$$

which is the definition of the solution of (5.4).

Remark 5.2. As in the previous section, we can show that

(1) We have

$$\lim_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon^3} \int_0^T \int_{\omega_\varepsilon} \varphi'_\varepsilon{}^2(x, t) dx dt = \frac{1}{3} \int_{\Sigma_0} \left(\frac{\partial \varphi'}{\partial \nu}(y, t) \right)^2 dy dt;$$

(2) We can reach initial data (y^0, y^1) in $H^{-1}(\Omega) \times (H^2(\Omega) \cap H_0^1(\Omega))'$.

Conclusion. We have proved that the exact controllability of the wave equation when the control is given by HUM and acts on the Dirichlet boundary condition and on Γ_0 can be obtained as the limit of the solutions of the exact controllability problems when the controls are given by HUM and act on an ε -neighborhood of Γ_0 whose lengths tend to 0.

REFERENCES

- [1] C. FABRE, *Equation des ondes avec second membre singulier et application à la contrôlabilité exacte*, Note C. R. Acad. Sci., 310 (1990), pp. 813–818.
- [2] ———, *An exact controllability problem related to the beams' equation*, in Proc. IFIP Conference of Irsee, 1990, to appear.
- [3] ———, *Quelques résultats de contrôlabilité exacte de l'équation de Schrödinger et application à l'équation des plaques vibrantes*, Note C. R. Acad. Sci., 312 (1991), pp. 61–66.
- [4] C. FABRE, AND J. P. PUEL, *Comportement au voisinage du bord des solutions de l'équation des ondes*, Note C. R. Acad. Sci., 310 (1990), pp. 621–625.
- [5] ———, *Behavior near the boundary for solutions of the wave equation*, to appear.
- [6] J. L. LIONS, *Exact controllability, stabilization and perturbations for distributed systems*, SIAM Rev., 30 (1988), pp. 1–68.
- [7] ———, *Contrôlabilité Exacte*, Collection RMA, Masson, Paris, 1988.
- [8] E. ZUAZUA, *Contrôlabilité exacte interne de l'équation des ondes*, Collection RMA, Masson, Paris, 1988.

A GENERALIZED ESTIMATE OF THE NUMBER OF ZEROS FOR SOLUTIONS OF A CLASS OF LINEAR DIFFERENTIAL EQUATIONS*

A. V. DMITRUK†

Abstract. For a class of triangular systems of N th-order linear ordinary differential equations with measurable coefficients, a simple proof of the following property is given: On every interval of an a priori prescribed length, which depends only on the norms of coefficients, the first component of every nontrivial solution of the system has at most $N - 1$ zeros.

Key words. differential equation, measurable coefficients, zeros of solutions

AMS(MOS) subject classifications. 34A30, 34C10

1. Introduction. It is well known that if a solution of the equation

$$(1) \quad x^{(N)} + \alpha_1(t)x^{(N-1)} + \cdots + \alpha_N(t)x = 0,$$

where $\alpha_1, \dots, \alpha_N$ are measurable and bounded real-valued functions of t , has infinitely many zeros on a certain time interval, then it vanishes identically on this interval (see, e.g., [1]). Moreover, for every positive real number A , there exists $T = T(N, A) > 0$ such that on every interval of length $\leq T$ and for every choice of coefficients α_k with $|\alpha_k(t)| \leq A$, $k = 1, \dots, N$, on this interval, every nontrivial solution of (1) has at most $N - 1$ zeros.

In [2] Sussmann proposes the following generalization of this property. He considers a triangular system of linear homogeneous equations of the form

$$(2) \quad \begin{aligned} \dot{x}_1 &= \alpha_{11}x_1 + \beta_1x_2, \\ \dot{x}_2 &= \alpha_{21}x_1 + \alpha_{22}x_2 + \beta_2x_3, \\ &\vdots \\ \dot{x}_k &= \alpha_{k1}x_1 + \cdots + \alpha_{kk}x_k + \beta_kx_{k+1}, \\ &\vdots \\ \dot{x}_N &= \alpha_{N1}x_1 + \cdots + \alpha_{NN}x_N, \end{aligned}$$

where α_{kj}, β_k are measurable and essentially bounded on a certain time interval $\Delta = [t_0, t_1]$, and x_k are absolutely continuous (the particular case, where $\alpha_{kj} = 0, \beta_k = 1$ for all $k < N$ and all $j \leq k$, corresponds to (1)), and the following result is stated.

LEMMA 1. *Let $A > B > 0$ be real numbers. Then there is a real number $T = T(N, A, B) > 0$ such that if the length of Δ is $\leq T$, and if*

$$(3) \quad |\alpha_{kj}(t)| \leq A, \quad B \leq \beta_k(t) \leq A \quad \text{a.e. on } \Delta,$$

then every nontrivial absolutely continuous solution of (2) has the property that its first component x_1 has $\leq N - 1$ zeros on Δ .

This property is used in [2] to estimate the number of switchings in an optimal control problem without singular regimes and has been appraised as the main technical point of that paper. The proof of Lemma 1 given in [2] is quite lengthy and very complicated. However, in our view, this result has intrinsic value, independent of the

* Received by the editors April 23, 1990; accepted for publication (in revised form) July 16, 1991.

† Central Economic-Mathematical Institute of the Academy of Sciences of the USSR, Moscow 117418, ul. Krasikova, 32, Russia.

particular application in [2], and a more direct proof is highly desirable. We give below a very simple and natural proof of Lemma 1.

(Note that since $\beta_k(t) \equiv B > 0$, every x_k , in a way, plays the role of the derivative of x_{k-1} ; so, qualitatively, system (2) is kindred to (1), and this is the crucial circumstance that allows us to spread the above-mentioned property to system (2).)

2. A new proof of Lemma 1. The key idea in our proof is to replace the triangular system of differential equations (2) by an equivalent, superdiagonal system. This idea is formalized in the next proposition.

PROPOSITION 1. *Let us consider the system*

$$(4) \quad \begin{aligned} \dot{x}_1 &= \alpha_{11}x_1 + \beta_1x_2, \\ \dot{x}_2 &= \alpha_{21}x_1 + \alpha_{22}x_2 + \beta_2x_3, \\ &\vdots \\ \dot{x}_k &= \alpha_{k1}x_1 + \cdots + \alpha_{kk}x_k + \beta_kx_{k+1}, \\ \dot{x}_{k+1} &= \alpha_{k+1,1}x_1 + \cdots + \alpha_{k+1,k+1}x_{k+1} + \beta_{k+1}x_{k+2}. \end{aligned}$$

Then, for every $A > 0$, there exists $l = l(k, A) > 0$ such that for every closed interval $\Delta = [t_0, t_1]$ of length $\leq l$ and for every choice of measurable coefficients satisfying the inequalities

$$(5) \quad |\alpha_{is}(t)| \leq A, \quad |\beta_i(t)| \leq A \quad \text{a.e. on } \Delta,$$

there exist absolutely continuous functions $\varphi_1(t), \dots, \varphi_k(t)$ of modulus ≤ 1 on Δ such that under the transformation

$$(6) \quad y_{k+1} = x_{k+1} + (\varphi_1x_1 + \cdots + \varphi_kx_k),$$

the last equation in (4) takes the form

$$(7) \quad \dot{y}_{k+1} = \lambda_{k+1}y_{k+1} + \beta_{k+1}x_{k+2}$$

(where $\lambda_{k+1} = \alpha_{k+1,k+1} + \beta_k\varphi_k$).

Proof. We compute \dot{y}_{k+1} using (6) and (4) and equate it to the right-hand side of (7). This yields the following system for φ_s :

$$(8) \quad \begin{aligned} \dot{\varphi}_s &= \beta_k\varphi_k\varphi_s + \alpha_{k+1,k+1}\varphi_s - \beta_{s-1}\varphi_{s-1} \\ &\quad - \sum_{i=s}^k \alpha_{is}\varphi_i - \alpha_{k+1,s}, \quad s = 1, \dots, k, \end{aligned}$$

where $\beta_0 = \varphi_0 = 0$ by convention.

This is a system of Riccati type. We will assign zero initial conditions as follows:

$$(9) \quad \varphi_s(t_0) = 0, \quad s = 1, \dots, k.$$

It is clear that, for every $A > 0$, there is $l = l(k, A) > 0$ such that on every interval of length $\leq l$, the solution of (8), (9) under condition (5) exists and satisfies $|\varphi_s(t)| \leq 1$. Thus transformation (6) with such φ_s , $s = 1, \dots, k$, yields the desired form (7). \square

This proposition allows us to simplify the last equation in (4) by eliminating the terms involving x_1, \dots, x_k without changing the coefficient β_{k+1} . Under this simplification, the coefficients of the k th equation are also transformed, but innocuously: the α_{ks} are transformed to $\alpha'_{ks} = \alpha_{ks} - \beta_k\varphi_s$ for $s = 1, \dots, k$, while β_k stays unchanged. The remaining equations are entirely unaffected.

This simplification procedure can be applied again to the k th equation in (4), then to the $(k-1)$ th equation in (4), and so on, up to the second equation. The first

equation requires no simplification. These repeated applications of the simplification procedure result in a system of the form

$$\begin{aligned}
 \dot{y}_1 &= \lambda_1 y_1 + \beta_1 y_2, \\
 \dot{y}_2 &= \lambda_2 y_2 + \beta_2 y_3, \\
 &\vdots \\
 \dot{y}_k &= \lambda_k y_k + \beta_k y_{k+1}, \\
 \dot{y}_{k+1} &= \lambda_{k+1} y_{k+1} + \beta_{k+1} y_{k+2},
 \end{aligned}
 \tag{10}$$

where $y_1 \equiv x_1$, $y_{k+2} \equiv x_{k+2}$, and the β_s , $s = 1, \dots, k+1$ are as before.

This transformation can be performed on any interval Δ with

$$|\Delta| \leq T(k, A) = \min_{s \leq k} l(s, 2A).$$

(We could also eliminate the coefficients λ_s by the substitution $v_s = y_s e^{-\Lambda_s(t)}$, where $\Lambda_s(t) = \int_{t_0}^t \lambda_s(\tau) d\tau$, so as to yield $\dot{v}_s = \exp(\Lambda_{s+1}(t) - \Lambda_s(t)) \beta_s v_{s+1}$.)

PROPOSITION 2 (a generalization of Rolle's theorem). *Let $y(t)$ be an absolutely continuous function on an opened interval $\omega = (t', t'')$ that satisfies the equation $\dot{y} = \lambda y + \beta z$ almost everywhere on ω , where $\lambda, \beta \in L_1(\omega)$, and $z(t)$ is continuous. Furthermore, suppose that $y(t') = y(t'') = 0$, $\beta(t) \geq 0$ on ω and*

$$\int_{\omega} \beta dt > 0. \tag{11}$$

Then z has at least one zero on ω .

Proof. Set $\Lambda(t) = \int_{t'}^t \lambda(\tau) d\tau$ and observe that

$$y(t'') = e^{\Lambda(t'')} \cdot \int_{\omega} e^{-\Lambda(\tau)} \beta(\tau) z(\tau) d\tau = 0. \tag{12}$$

If z has no zeros on ω , then by continuity it has a constant sign, e.g., $z > 0$ on ω . Then, from (11) and the nonnegativity of β , we see that the integral in (12) is positive, which is a contradiction.

Now we return to the original system (2).

LEMMA 2. *For every $A > 0$, there exists $T = T(N, A) > 0$ such that for every interval Δ with length $|\Delta| \leq T$, for every choice of measurable coefficients, which satisfy on Δ inequalities (5) and are such that $\beta_k(t) \geq 0$ almost everywhere on Δ , $k = 1, \dots, N$, and for every open interval $\omega \subset \Delta$*

$$\int_{\omega} \beta_k dt > 0, \tag{13}$$

the following property holds: If (x_1, \dots, x_N) is a solution of (2) and x_1 has at least N distinct zeros on Δ , then all the x_k for $k = 1, \dots, N$ are identically zero.

Proof. It was proved above that on every interval Δ with $|\Delta| \leq T(N, A)$, system (2) can be transformed to the form

$$\begin{aligned}
 \dot{y}_1 &= \lambda_1 y_1 + \beta_1 y_2, \\
 \dot{y}_2 &= \lambda_2 y_2 + \beta_2 y_3, \\
 &\vdots \\
 \dot{y}_{N-1} &= \lambda_{N-1} y_{N-1} + \beta_{N-1} y_N, \\
 \dot{y}_N &= \lambda_N y_N,
 \end{aligned}
 \tag{14}$$

where $\beta_N \equiv 0$ and $y_1 \equiv x_1$.

Suppose that $x_1 \equiv y_1$ has at least N distinct zeros on Δ . By Proposition 2 and condition (13), between every two successive zeros of y_1 , there is a zero of y_2 , and so y_2 has at least $N-1$ distinct zeros on Δ . Between every two successive zeros of y_2 there is a zero of y_3 , and so on. Coming to y_N , we infer that it has at least one zero on Δ . Then, by the homogeneity of the last equation in (14), we get $y_N \equiv 0$. Then, however, the $(N-1)$ th equation is also homogeneous, and since y_{N-1} has zeros on Δ , then $y_{N-1} \equiv 0$, and so on. We deduce that $y_k \equiv 0$ for each $k = 1, \dots, N$, which clearly implies that $x_k \equiv 0$ for each $k = 1, \dots, N$. \square

Remark. This lemma is slightly stronger than Lemma 1, since the conditions on β_k are slightly weaker than those in Lemma 1—namely, it is not assumed that β_k has a positive lower bound. A further relaxation of the requirements on coefficients of (2) is possible, as we show in the next section.

3. A generalization of Lemma 1. In this section we show that the uniform constraints (5) can be replaced by integral constraints. Let us denote

$$C_\Delta = \max \left\{ \max_{k,j} \int_\Delta |\alpha_{kj}| dt, \max_k \int_\Delta \beta_k dt \right\}.$$

LEMMA 3. Let Δ be a closed interval, let $\alpha_{kj}, \beta_k \in L_1(\Delta)$, $k = 1, \dots, N, j = 1, \dots, k$ be such that the functions β_k are nonnegative and satisfy (13), and suppose that

$$(15) \quad C_\Delta \leq \max_{H \geq 0} \frac{H}{H^2 + 2NH + 1}.$$

Then system (2) has the following property: If x_1 has at least N zeros on Δ , then all the components x_k , $k = 1, \dots, N$, are identically zero on Δ .

(We do not need to compute the explicit value of the maximum (15); the positiveness of this maximum, which is evident, is sufficient for our purposes.)

Proof. It is sufficient to show that under condition (15) the transformation (6) can still be performed, i.e., that for every $k < N$, system (8), (9) has a solution defined on the whole interval Δ . The local existence of such a solution is obvious; the only obstacle may be excessively rapid growth of the solution. We denote

$$p(t) = \max_{1 \leq s \leq k} |\varphi_s(t)|.$$

Due to (15) there exists $H > 0$ such that

$$C_\Delta(H^2 + NH + 1) \leq H.$$

We claim that this inequality implies that $p(t) < H$ on the whole closed interval $\Delta = [t_0, t_1]$.

Suppose not, and let $t_* \in \Delta$ be the first point where $p(t) = H$. It is clear that $t_* > t_0$, since $p(t_0) = 0$. Then, on (t_0, t_*) , we get $p < H$ and, from (8),

$$|\dot{\varphi}_s| < \beta_k H^2 + \left(|\alpha_{k+1,k+1}| + |\beta_{s-1}| + \sum_{i=s}^k |\alpha_{is}| \right) H + |\alpha_{k+1,s}|,$$

$$s = 1, \dots, k,$$

which implies that

$$p(t_*) < C_\Delta(H^2 + NH + 1) \leq H.$$

This, however, contradicts the assumption that $p(t_*) = H$, and the claim is established.

Thus, on the whole interval Δ and for every $k < N$, system (8) has a solution with the property $p(t) < H$. We thus can obtain system (14) as before and repeat the argument of Lemma 2, thereby proving Lemma 3. \square

It is obvious that lemma 2 follows from Lemma 3. For given $A > 0$, we can take any T with

$$AT \leq \max_{H \geq 0} \frac{H}{H^2 + NH + 1}.$$

Due to (5), we have $C_\Delta \leq AT$, and so (15) holds.

REFERENCES

- [1] P. HARTMAN, *Ordinary Differential Equations*, John Wiley, New York, 1964.
- [2] H. J. SUSSMANN, *A bang-bang theorem with bounds on the number of swithings*, SIAM J. Control Optim., 17 (1979), pp. 629-651.

A DIRICHLET BOUNDARY CONTROL PROBLEM FOR THE STRONGLY DAMPED WAVE EQUATION*

FRANCESCA BUCCI†

Abstract. A boundary control problem is considered for the strongly damped wave equation, and it is solved by dynamic programming arguments.

Key words. boundary control, Riccati equation, dynamic programming

AMS(MOS) subject classifications. 49, 49C20, 49B22

1. Introduction.

1.1. Statement of the problem and literature. Let $\Omega \subset \mathbb{R}^n$ be an open bounded set with smooth boundary $\partial\Omega$, and let $T > 0$ be fixed.

We are concerned with a boundary control problem for the strongly damped wave equation

$$\begin{aligned} y_{tt}(t, x) &= \Delta y(t, x) + c\Delta y_t(t, x) & (t, x) \in]0, T[\times \Omega; \\ (1.1) \quad y(0, x) &= y_0(x), \quad y_t(0, x) = z_0(x) & x \in \Omega; \\ y(t, x) &= u(t, x) & (t, x) \in]0, T[\times \partial\Omega, \end{aligned}$$

where c is a positive constant; $y_0, z_0 \in L^2(\Omega)$; and we take u in $W^{1,2}(0, T; L^2(\partial\Omega))$.

Physical motivation for studying (1.1) arises from problems that may occur in the study of flexible structures in a bounded domain, controlled on the boundary through a Dirichlet boundary condition.

In recent years, boundary control problems have become of interest in optimal control theory. Flandoli [2], [3] and Lasiecka and Triggiani [4] study a general abstract class of dynamic that covers parabolic-like problems, namely, not only heat/diffusion equations, but also wave or plate equations with structural damping. In their works, they assume, as usual, that controls u belong to $L^2(0, T; L^2(\partial\Omega))$.

In [5] Lasiecka and Triggiani give several examples of partial differential equations, with boundary or point control, which can be reduced to that abstract model. Nevertheless, to the knowledge of the author, (1.1) has not been explicitly treated in relation to optimal control problems.

In this paper, following the original idea of Balakrishnan for parabolic equations (see [1]), we derive a solution formula for (1.1) in the product space $H = L^2(\Omega) \times L^2(\Omega)$. This formula yields the couple (y, y_t) in terms of the time derivative of the control u_t .

Since we want to solve the control problem using dynamic programming techniques, we must work in the product space H , and we would expect that (y, y_t) belongs to $L^2(0, T; H)$. Therefore, due to the low regularity of the solutions to (1.1) under the assumption $u \in L^2(0, T; L^2(\partial\Omega))$, we take $u \in W^{1,2}(0, T; L^2(\partial\Omega))$.

* Received by the editors August 15, 1990; accepted for publication (in revised form) June 26, 1991.

† Dipartimento di Matematica, Via Buonarroti 2, 56127 Pisa, Italy.

Consistent with this choice, we consider the problem of minimizing the cost functional

$$\begin{aligned}
 J(u) = & \int_0^T dt \int_{\Omega} \{ |(C_1 y(t, \cdot))(x)|^2 + |(C_2 y_t(t, \cdot))(x)|^2 \} dx \\
 (1.2) \quad & + \int_0^T dt \int_{\partial\Omega} \{ |u(t, x)|^2 + |u_t(t, x)|^2 \} d\sigma \\
 & + \int_{\Omega} \{ |(\Gamma_1 y(T, \cdot))(x)|^2 + |(\Gamma_2 y_t(T, \cdot))(x)|^2 \} dx
 \end{aligned}$$

overall u in $W^{1,2}(0, T; L^2(\partial\Omega))$, where $C_i, \Gamma_i \in \mathcal{L}(L^2(\Omega))$, $i = 1, 2$, Γ_i are selfadjoint, and y is subject to the partial differential equation (1.1).

The purpose of § 2 of this paper is to show that it is possible to reformulate problem (1.1), (1.2) into a standard quadratic control problem. This goal is achieved by introducing suitable states and controls, namely, setting $W = (y \cdot y_t, u)$, $v = u'$.

Section 3 is devoted to showing that the theory developed in [3] can be applied to the new control problem, provided that Γ_2 belong to $\mathcal{L}(L^2(\Omega), H_0^{2\beta}(\Omega))$ for some $\beta \in (\frac{1}{4}, \frac{1}{2})$.

1.2. Notation. Let X and Y be two Hilbert spaces. We denote norms and inner products with $|\cdot|$ and $\langle \cdot, \cdot \rangle$, respectively.

We represent with $\mathcal{L}(X, Y)$ ($\mathcal{L}(X)$ if $X = Y$), $\Sigma(X)$, $\Sigma^+(X)$ the space of all bounded linear operators from X to Y , the space of all bounded selfadjoint operators in X , and the subset of $\Sigma(X)$ of nonnegative definite operators, respectively.

If T is a linear operator (generally unbounded) from X to Y , we denote its domain with $D(T)$ and its adjoint by T^* .

Moreover, we denote by $\rho(T)$ the resolvent set of T , by $\sigma(T)$ the spectrum of T , and by $R(\lambda, T) = (\lambda - T)^{-1}$ the resolvent operator, respectively. We set $\omega_T = \sup \{ \operatorname{Re} \lambda \mid \lambda \in \sigma(T) \}$.

If T generates a C_0 -semigroup $G(t)$ on X , we set $G(t) = e^{tT}$.

2. The abstract setting. Let $\Omega \subset \mathbb{R}^n$ be an open bounded set with smooth boundary $\partial\Omega$, and let $T > 0$ be fixed. We study in $(0, T) \times \Omega$ the optimal control problem (1.1), (1.2).

We consider the Dirichlet realization of the Laplace operator in $L^2(\Omega)$, defined by $Ay = \Delta y$ for any $y \in D(A) = H^2(\Omega) \cap H_0^1(\Omega)$, and we denote by D the Dirichlet mapping from $L^2(\partial\Omega)$ to $L^2(\Omega)$, defined by $Dv = w$, where

$$\Delta w = 0 \quad \text{in } \Omega,$$

$$w(x) = v(x) \quad x \in \partial\Omega.$$

As is proved in [6], $D \in \mathcal{L}(L^2(\partial\Omega), H^{1/2}(\Omega))$.

Moreover, we introduce the Hilbert spaces $H = L^2(\Omega) \times L^2(\Omega)$, $U = L^2(\partial\Omega)$ and define the linear operator \mathcal{A} in the product space H as

$$\begin{aligned}
 (2.1) \quad \mathcal{A} \begin{pmatrix} y \\ z \end{pmatrix} &= \begin{pmatrix} 0 & I \\ A & cA \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} z \\ A(y + cz) \end{pmatrix}, \\
 D(\mathcal{A}) &= \left\{ \begin{pmatrix} y \\ z \end{pmatrix} \in H \mid y + cz \in D(A) \right\}.
 \end{aligned}$$

It is well known that \mathcal{A} is the infinitesimal generator of an analytic semigroup $e^{t\mathcal{A}}$ on H of negative type.

For simplicity, we first assume that $u(0, \cdot) = 0$ and consider the problem of minimizing (1.2) over the class of controls

$$W_0^{1,2}(0, T; L^2(\partial\Omega)) = \{u \in W^{1,2}(0, T; L^2(\partial\Omega)) \mid u(0, \cdot) = 0\}.$$

Following a standard technique introduced by Balakrishnan (see [1]), we can reduce problem (1.1) to a homogeneous boundary problem. Then it is easy to check that the solution $Y = (y, y_t)$ to (1.1) satisfies

$$(2.2) \quad Y(t) = e^{t\mathcal{A}} \begin{pmatrix} y_0 \\ z_0 \end{pmatrix} - \mathcal{A} \int_0^t e^{(t-s)\mathcal{A}} Fu'(s) ds + Eu(t),$$

where $Y(t) = Y(t, \cdot)$, $u(t) = u(t, \cdot)$, and E, F are the linear operators in $\mathcal{L}(U, H)$, defined by

$$Eu = \begin{pmatrix} Du \\ 0 \end{pmatrix}, \quad Fu = \begin{pmatrix} 0 \\ Du \end{pmatrix},$$

respectively.

Remark 2.1. If we apply

$$\mathcal{A}^{-1} = \begin{pmatrix} -cI & A^{-1} \\ I & 0 \end{pmatrix}$$

to (2.2) and integrate by parts in t , we obtain that

$$\begin{pmatrix} -cy(t) + A^{-1}y'(t) \\ y(t) \end{pmatrix} = \mathcal{A}^{-1} e^{t\mathcal{A}} \begin{pmatrix} y_0 \\ z_0 \end{pmatrix} - \mathcal{A} \int_0^t e^{(t-s)\mathcal{A}} Fu(s) ds - cEu(t),$$

which easily yields regularity of solutions to (1.1) in terms of the regularity of the control u . We stress that, if $u \in L^2(0, T; L^2(\partial\Omega))$, then we only have $A^{-1}y'(t) \in L^2(0, T; H)$.

Therefore, because we want to use dynamic programming arguments, we cannot weaken the assumption $u \in W^{1,2}(0, T; L^2(\partial\Omega))$.

The cost functional can be written as

$$(2.3) \quad J(u) = \int_0^T \{|CY(s)|^2 + |u(s)|^2 + |u'(s)|^2\} ds + \langle P_0 Y(T), Y(T) \rangle,$$

where

$$C = \begin{pmatrix} C_1 & 0 \\ 0 & C_2 \end{pmatrix}, \quad P_0 = \begin{pmatrix} \Gamma_1^2 & 0 \\ 0 & \Gamma_2^2 \end{pmatrix},$$

and it is clear that $C \in \mathcal{L}(H)$, $P_0 \in \Sigma^+(H)$.

Now note the control u as an auxiliary component of the state and define u' as a new control. More precisely, set

$$(2.4) \quad u' = v, \quad W = \begin{pmatrix} Y \\ u \end{pmatrix}$$

and introduce a new states space $\bar{H} = H \times U$, while we set $\bar{U} = U$.

From (2.2), (2.4) it is rather easy to derive a semigroup formula to be satisfied by W in \bar{H} . To do that, we need the following lemma.

LEMMA 2.2. Let $G: [0, +\infty) \rightarrow \mathcal{L}(\bar{H})$ be defined by

$$(2.5) \quad t \rightarrow \begin{pmatrix} e^{t\mathcal{A}} & (I - e^{t\mathcal{A}})E \\ 0 & I \end{pmatrix},$$

with \mathcal{A} given by (2.1).

Then G is an analytic semigroup on \bar{H} of type $=0$, and its generator \mathcal{B} is defined by

$$(2.6) \quad \begin{aligned} D(\mathcal{B}) &= \left\{ \begin{pmatrix} Y \\ u \end{pmatrix} \in \bar{H} \mid Y - Eu \in D(\mathcal{A}) \right\}, \\ \mathcal{B} \begin{pmatrix} Y \\ u \end{pmatrix} &= \begin{pmatrix} \mathcal{A}(Y - Eu) \\ 0 \end{pmatrix}. \end{aligned}$$

Moreover, $\rho(\mathcal{B}) = \{\lambda \in \mathbb{C} \mid \lambda \in \rho(\mathcal{A}), \lambda \neq 0\}$, and, for all $\lambda \in \rho(\mathcal{B})$, we have that

$$(2.7) \quad R(\lambda, \mathcal{B}) = \begin{pmatrix} R(\lambda, \mathcal{A}) & -(1/\lambda)\mathcal{A}R(\lambda, \mathcal{A})E \\ 0 & (1/\lambda)I \end{pmatrix}.$$

Proof. We can easily check that G is a strongly continuous semigroup on \bar{H} . We now characterize the generator \mathcal{B} of $G(t)$.

Let $\begin{pmatrix} Y \\ u \end{pmatrix} \in \bar{H}$, $t > 0$. We write

$$\begin{aligned} \frac{1}{t}(G(t) - I) \begin{pmatrix} Y \\ u \end{pmatrix} &= \frac{1}{t} \begin{pmatrix} e^{t\mathcal{A}} - I & (I - e^{t\mathcal{A}})E \\ 0 & 0 \end{pmatrix} \begin{pmatrix} Y \\ u \end{pmatrix} \\ &= \frac{1}{t} \begin{pmatrix} (e^{t\mathcal{A}} - I)(Y - Eu) \\ 0 \end{pmatrix}. \end{aligned}$$

Therefore the limit $\lim_{t \rightarrow 0} (1/t)(G(t) - I)\begin{pmatrix} Y \\ u \end{pmatrix}$ exists if and only if there exists $\lim_{t \rightarrow 0} (1/t)(e^{t\mathcal{A}} - I)(Y - Eu)$, that is, by definition, if $Y - Eu \in D(\mathcal{A})$.

In conclusion,

$$D(\mathcal{B}) = \left\{ \begin{pmatrix} Y \\ u \end{pmatrix} \in \bar{H} \mid Y - Eu \in D(\mathcal{A}) \right\}$$

and, for all $\begin{pmatrix} Y \\ u \end{pmatrix} \in D(\mathcal{B})$,

$$\mathcal{B} \begin{pmatrix} Y \\ u \end{pmatrix} = \begin{pmatrix} \mathcal{A}(Y - Eu) \\ 0 \end{pmatrix},$$

and (2.6) holds true.

Also, formula (2.7) can be easily verified.

To show that $G(t) = e^{t\mathcal{B}}$ is an analytic semigroup on \bar{H} , we observe that, if $\omega > 0$, we have that

$$\rho(\mathcal{B}) \supset \{\lambda \in \mathbb{C} \mid \operatorname{Re} \lambda > \omega\},$$

and, since $e^{t\mathcal{A}}$ is an analytic semigroup of negative type, from (2.7) we easily deduce the bound

$$|R(\lambda, \mathcal{B})| \leq \frac{M}{|\lambda - \omega|} \quad \operatorname{Re} \lambda > \omega. \quad \square$$

Remark 2.3. By using (2.6) and definition (2.1) of \mathcal{A} , we can write more explicitly

$$\begin{aligned} D(\mathcal{B}) &= \left\{ \begin{pmatrix} y \\ z \\ u \end{pmatrix} \in \bar{H} \mid y + cz - Du \in D(A) \right\}, \\ \mathcal{B} \begin{pmatrix} y \\ z \\ u \end{pmatrix} &= \begin{pmatrix} z \\ A(y + cz - Du) \\ 0 \end{pmatrix}. \end{aligned}$$

By using the operators defined in Lemma 2.2, we can finally obtain the following theorem.

THEOREM 2.4. *Let Y be as in (2.2), W and v defined by (2.4). Then $W(t)$ satisfies*

$$(2.8) \quad W(t) = e^{t\mathcal{B}} W_0 + (I - \mathcal{B}) \int_0^t e^{(t-s)\mathcal{B}} \mathcal{G}v(s) ds,$$

where $e^{t\mathcal{B}}$, \mathcal{B} are given by (2.5), (2.6), respectively; \mathcal{G} is the linear bounded operator from \bar{U} to \bar{H} defined by

$$(2.9) \quad \mathcal{G}v = \begin{pmatrix} Ev - \mathcal{A}(I - \mathcal{A})^{-1}Fv \\ v \end{pmatrix};$$

and $W_0 = (y_0, z_0, 0)^T$.

Proof. Formula (2.8) is proved by a short verification substituting (2.9) and $W_0 = (y_0, z_0, 0)^T$ into the second member of (2.8) and considering (2.5) and (2.6). \square

In conclusion, the control problem (2.2), (2.3) can be reduced, in the abstract spaces \bar{H} , \bar{U} , to the problem of minimizing the quadratic functional

$$(2.10) \quad J(v) = \int_0^T (|\bar{C}W(s)|^2 + |v(s)|^2) ds + \langle \bar{P}_0 W(T), W(T) \rangle,$$

over all $v \in L^2(0, T; \bar{U})$, where

$$\bar{C} = \begin{pmatrix} C_1 & 0 & 0 \\ 0 & C_2 & 0 \\ 0 & 0 & I \end{pmatrix}, \quad \bar{P}_0 = \begin{pmatrix} \Gamma_1^2 & 0 & 0 \\ 0 & \Gamma_2^2 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

and W is subject to (2.8).

Suppose now that \mathcal{B} , \mathcal{G} , \bar{C} , \bar{P}_0 satisfy all conditions assumed by Flandoli in [3] to show the existence and uniqueness of the solutions to Riccati equation associated with problem (2.8)–(2.10). Obviously, once we have obtained the optimal control $v^* \in L^2(0, T; \bar{U})$ for problem (2.8)–(2.10), the optimal control u^* for the original problem (1.1), (1.2) is given by $u^*(t) = \int_0^t v^*(s) ds$.

Remark 2.5. Until now, we have supposed that $u(0) = 0$. Otherwise, we can proceed as follows. We first assume that $u(0) = u_0 \in U$ is fixed, and we derive the solution formula for (1.1) in H as follows:

$$(2.11) \quad Y(t) = e^{t\mathcal{A}} \begin{pmatrix} y_0 - Du_0 \\ z_0 \end{pmatrix} - \mathcal{A} \int_0^t e^{(t-s)\mathcal{A}} Fu'(s) ds + Eu(t).$$

By using the same method described in the case where $u_0 = 0$, we reduce the problem of minimizing (2.3) over the class of controls $u \in W^{1,2}(0, T; U)$ such that $u(0) = u_0$ (where Y is subject to (2.11)) to problem (2.8)–(2.10), with $W_0 = (y_0, z_0, u_0)$.

If the theory developed in [3] applies to (2.8)–(2.10), then the Riccati feedback synthesis yields the optimal value $J(v^*) = \langle P(T)W_0, W_0 \rangle$, where P is the solution to the Riccati equation associated with (2.8)–(2.10). This is a quadratic form with respect to u_0 . Thus, to solve the original control problem in $W^{1,2}(0, T; U)$, it remains to minimize $J(v^*)$ with respect to u_0 .

3. Solution of the control problem. We want to check hypotheses assumed in [3] to solve problem (2.8)–(2.10). We can immediately see that $\bar{C} \in \mathcal{L}(\bar{H})$, $\bar{P}_0 \in \Sigma^+(\bar{H})$. As a consequence of Lemma 2.2, we also know that \mathcal{B} generates an analytic semigroup of type less than 1.

Therefore it remains to show that

$$(3.1) \quad \exists 0 < \alpha < 1 \quad \text{such that } \mathcal{G} \in \mathcal{L}(\bar{U}, D((I - \mathcal{B})^\alpha))$$

and that, under suitable assumptions on Γ_i ,

$$(3.2) \quad \exists \beta \in (\tfrac{1}{2} - \alpha, \tfrac{1}{2}) \quad \text{such that } (I - \mathcal{B}^*)^\beta \sqrt{\bar{P}_0} \in \mathcal{L}(\bar{H}).$$

To prove the validity of (3.1), we give a characterization of the interpolation spaces $D_{\mathcal{B}}(\theta, 2)$, for any $\theta \in (0, 1)$. (As to relations between interpolation spaces and domains of fractional powers of linear operators, see [7, §§ 1.13–1.15], and the references contained therein.)

We start by showing the following lemma.

LEMMA 3.1. *For any $\theta \in (0, 1)$*

$$(3.3) \quad D_{\mathcal{B}}(\theta, 2) = \left\{ \begin{pmatrix} Y \\ u \end{pmatrix} \in \bar{H} \mid Y - Eu \in D_{\mathcal{A}}(\theta, 2) \right\},$$

and the norm

$$(3.4) \quad \begin{pmatrix} Y \\ u \end{pmatrix} \rightarrow \left\| \begin{pmatrix} Y \\ u \end{pmatrix} \right\|_{\bar{H}} + \|Y - Eu\|_{D_{\mathcal{A}}(\theta, 2)}$$

is equivalent to the norm of $D_{\mathcal{B}}(\theta, 2)$.

Proof. We use the well-known characterization [7, § 1.14], below:

$$D_{\mathcal{B}}(\theta, p) = \{ W \in \bar{H} : t \rightarrow \|t^\theta \mathcal{B}R(t, \mathcal{B})W\| \in L_*^p(a, +\infty) \}$$

with norm

$$(3.5) \quad \|W\|_{D_{\mathcal{B}}(\theta, p)} = \|W\|_{\bar{H}} + \|t^\theta \mathcal{B}R(t, \mathcal{B})W\|_{L_*^p(a, +\infty)},$$

where $a \geq \max(1, \omega_{\mathcal{B}})$, and $f \in L_*^p(a, +\infty)$ if

$$\int_a^{+\infty} |f(t)|^p \frac{dt}{t} < +\infty.$$

Let $\theta \in (0, 1)$, $\begin{pmatrix} Y \\ u \end{pmatrix} \in \bar{H}$, $t \geq a$. By representation (2.7) of the resolvent $R(t, \mathcal{B})$ in terms of $R(t, \mathcal{A})$, it follows that

$$(3.6) \quad \mathcal{B}R(t, \mathcal{B}) \begin{pmatrix} Y \\ u \end{pmatrix} = \begin{pmatrix} \mathcal{A}R(t, \mathcal{A})(Y - Eu) \\ 0 \end{pmatrix}.$$

Therefore

$$t \rightarrow \left\| t^\theta \mathcal{B}R(t, \mathcal{B}) \begin{pmatrix} Y \\ u \end{pmatrix} \right\| \in L_*^2(a, +\infty)$$

if and only if

$$t \rightarrow \|t^\theta \mathcal{A}R(t, \mathcal{A})(Y - Eu)\| \in L_*^2(a, +\infty),$$

and (3.3) holds true. The equivalence of the norms (3.4), (3.5) is again a consequence of (3.6). \square

Arguing as in Lemma 3.1, by means of the representation of the resolvent $R(t, \mathcal{A})$ in terms of $R(t^2/(ct+1); A)$, we can easily deduce the next lemma.

LEMMA 3.2. *For any $\theta \in (0, 1)$,*

$$D_{\mathcal{A}}(\theta, 2) = \left\{ \begin{pmatrix} y \\ z \end{pmatrix} \in H \mid y + cz \in D_A(\theta, 2) \right\},$$

and the norm

$$\begin{pmatrix} y \\ z \end{pmatrix} \rightarrow \left\| \begin{pmatrix} y \\ z \end{pmatrix} \right\|_H + \|y + cz\|_{D_A(\theta, 2)}$$

is equivalent to the norm of $D_{\mathcal{A}}(\theta, 2)$.

COROLLARY 3.3. For any $\theta \in (0, 1)$, we have that

$$D_{\mathcal{B}}(\theta, 2) = \left\{ \begin{pmatrix} y \\ z \\ u \end{pmatrix} \in \bar{H} \mid y + cz - Du \in D_A(\theta, 2) \right\},$$

and the norm

$$\begin{pmatrix} y \\ z \\ u \end{pmatrix} \rightarrow \left\| \begin{pmatrix} y \\ z \\ u \end{pmatrix} \right\|_{\bar{H}} + \|y + cz - Du\|_{D_A(\theta, 2)}$$

is equivalent to the norm of $D_{\mathcal{B}}(\theta, 2)$.

We are now able to verify condition (3.1).

PROPOSITION 3.4. Let \mathcal{G} be as in (2.9). Then there exists $\theta \in (0, \frac{1}{4})$ such that $\mathcal{G} \in \mathcal{L}(\bar{U}, D((I - \mathcal{B})^\theta))$.

Proof. As a consequence of the inclusion

$$(3.7) \quad D_{\mathcal{B}}(\theta + \varepsilon, 2) \hookrightarrow D((I - \mathcal{B})^\theta),$$

which holds for any $\theta \in (0, 1)$, $\varepsilon > 0$, it is sufficient to show that there exists $\theta \in (0, \frac{1}{4})$ such that $\mathcal{G} \in \mathcal{L}(\bar{U}, D_{\mathcal{B}}(\theta, 2))$.

Let $v \in \bar{U}$. From Lemma 3.1, $\mathcal{G}v \in D_{\mathcal{B}}(\theta, 2)$ for some $\theta \in (0, 1)$ if and only if $R(1, \mathcal{A})Fv - Fv \in D_{\mathcal{A}}(\theta, 2)$ for the same θ .

Since $D \in \mathcal{L}(U, D_A(\theta, 2))$ for any $\theta \in (0, \frac{1}{4})$ [6], conclusion follows easily from Lemma 3.2. \square

It remains to check (3.2).

Let θ be as in Proposition 3.4. It is sufficient to show the existence of $\beta \in (\frac{1}{2} - \theta, \frac{1}{2})$ such that

$$(3.8) \quad \sqrt{\bar{P}_0} \in \mathcal{L}(\bar{H}, D_{\mathcal{B}^*}(\beta, 2)),$$

where \mathcal{B}^* is the adjoint of \mathcal{B} . After that, again as a consequence of (3.7)—which also holds true for \mathcal{B}^* —and by the closed graph theorem, we obtain that

$$\exists \beta \in (\frac{1}{2} - \theta, \frac{1}{2}) \quad \text{such that } (I - \mathcal{B}^*)^\beta \sqrt{\bar{P}_0} \in \mathcal{L}(\bar{H}).$$

By using the same arguments as in Lemmas 3.1 and 3.2, we can easily deduce the next lemma.

LEMMA 3.5. For any $\theta \in (0, 1)$,

$$(3.9) \quad D_{\mathcal{B}^*}(\theta, 2) = \left\{ \begin{pmatrix} y \\ z \\ u \end{pmatrix} \in \bar{H} : z \in D_A(\theta, 2) \right\},$$

and the norm

$$\begin{pmatrix} y \\ z \\ u \end{pmatrix} \rightarrow \left\| \begin{pmatrix} y \\ z \\ u \end{pmatrix} \right\|_{\bar{H}} + \|z\|_{D_A(\theta, 2)}$$

is equivalent to the norm of $D_{\mathcal{B}^*}(\theta, 2)$.

Assume now that

$$(3.10) \quad \exists \beta \in (\tfrac{1}{2} - \theta, \tfrac{1}{2}) \quad \text{such that } \Gamma_2 \in \mathcal{L}(L^2(\Omega), D_A(\beta, 2)).$$

Then we have the following proposition.

PROPOSITION 3.6. *There exists $\beta \in (\tfrac{1}{2} - \theta, \tfrac{1}{2})$ such that*

$$(I - \mathcal{B}^*)^\beta \sqrt{\bar{P}_0} \in \mathcal{L}(\bar{H}).$$

Proof. Let

$$\begin{pmatrix} y \\ z \\ u \end{pmatrix} \in \bar{H}.$$

Then

$$\sqrt{\bar{P}_0} \begin{pmatrix} y \\ z \\ u \end{pmatrix} = \begin{pmatrix} \Gamma_1 y \\ \Gamma_2 z \\ 0 \end{pmatrix}.$$

By hypothesis (3.10) on Γ_2 , and from (3.9), there exists $\beta \in (\tfrac{1}{2} - \theta, \tfrac{1}{2})$ such that

$$\sqrt{\bar{P}_0} \begin{pmatrix} y \\ z \\ u \end{pmatrix} \in D_{\mathcal{B}^*}(\beta, 2).$$

Moreover, as a consequence of Lemma 3.5, we can write

$$\left\| \sqrt{\bar{P}_0} \begin{pmatrix} y \\ z \\ u \end{pmatrix} \right\|_{D_{\mathcal{B}^*}(\beta, 2)} = \left\| \begin{pmatrix} \Gamma_1 y \\ \Gamma_2 z \\ 0 \end{pmatrix} \right\|_{\bar{H}} + \|\Gamma_2 z\|_{D_A(\beta, 2)},$$

and, again by (3.10), we deduce the bound

$$\left\| \sqrt{\bar{P}_0} \begin{pmatrix} y \\ z \\ u \end{pmatrix} \right\|_{D_{\mathcal{B}^*}(\beta, 2)} \leq \text{const} \left\| \begin{pmatrix} y \\ z \\ u \end{pmatrix} \right\|_{\bar{H}}.$$

Thus (3.8) holds true, and Proposition 3.6 is proved. \square

Now, following [3], we can solve the Riccati equation associated with (2.8)–(2.10) and conclude by using dynamic programming that, for every $y_0, z_0 \in L^2(\Omega)$, there exists a unique feedback optimal control v^* for (2.8)–(2.10).

At this point, we can interpret the Riccati feedback synthesis of problem (2.8)–(2.10) in terms of the original control problem (1.1), (1.2).

Therefore we can finally state the following theorem.

THEOREM 3.7. *If $C_i \in \mathcal{L}(L^2(\Omega))$, $\Gamma_i \in \Sigma(L^2(\Omega))$, $i = 1, 2$, $\Gamma_2 \in \mathcal{L}(L^2(\Omega), H_0^{2\beta}(\Omega))$ for some $\beta \in (\tfrac{1}{4}, \tfrac{1}{2})$, then, for every $y_0, z_0 \in L^2(\Omega)$, there exists a unique feedback optimal control u^* for problem (1.1), (1.2) in $W_0^{1,2}(0, T; L^2(\partial\Omega))$.*

REFERENCES

- [1] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Springer-Verlag, New York, 1976.
- [2] F. FLANDOLI, *Riccati equation arising in a boundary control problem with distributed parameters*, SIAM J. Control Optim., 22 (1984) pp. 76–86.

- [3] F. FLANDOLI, *On the direct solutions of Riccati equations arising in boundary control theory*, Ann. Mat. Pura Appl., to appear.
- [4] I. LASIECKA AND R. TRIGGIANI, *Dirichlet boundary control problem for parabolic equations with quadratic cost: Analyticity and Riccati's feedback synthesis*, SIAM J. Control Optim., 21 (1983), pp. 41–68.
- [5] ———, *Differential and Algebraic Riccati Equations with Applications to Boundary/Point Control Problems: Continuous Theory and Approximation Theory*, Lecture Notes in Control and Inform. Sci., 164, Springer-Verlag, Berlin, New York, 1991.
- [6] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, Springer-Verlag, New York, 1971.
- [7] H. TRIEBEL, *Interpolation Theory, Function Spaces, Differential Operators*, VEB Deutschen Verlag der Wissenschaften, Berlin, 1978.

CONTROLLABILITY OF l^2 -SYSTEMS*

FABIO FAGNANI[†] AND JAN C. WILLEMS[‡]

Abstract. This paper is devoted to an investigation of controllability and almost controllability of l^2 -systems. These concepts are defined in terms of the possibility of steering one system trajectory to another. It is proved that a controllable l^2 -system always has finite memory. The main result on almost controllability states that this is equivalent to the existence of a scattering representation. The paper ends with an investigation of the relation of almost controllability and state representations.

Key words. controllability, l^2 -systems, state representation, linear systems

AMS(MOS) subject classification. 93B05

1. Introduction. Controllability has played an instrumental role in the development of control theory during the past three decades and is now a fundamental concept in mathematical systems theory. It plays a central role in control synthesis questions, related to the very possibility of exerting effective control. As such, it enters as a crucial “existence” condition in many engineering-type questions, such as stabilization and optimal control.

The notion of controllability is usually introduced for state space representations [1], [10], where it refers to the possibility of transferring the state from an initial to a terminal value. For finite-dimensional, linear, time-invariant systems, controllability then implies that any initial state can be exactly transferred to any terminal state in finite time. For nonlinear systems, we must often be satisfied with a local version of this property. For infinite-dimensional systems, on the other hand, approximate controllability and/or variations in which we allow the transfer time to go to infinity have proved to be more relevant. In fact, the question of which, and in what sense, systems described by partial differential equations are controllable is far from settled (see [9]).

Recently, a notion of controllability was introduced, where it becomes an intrinsic property of a dynamical system, and not just of a state space representation [12], [13]. The basic idea is to call a system controllable if an arbitrary past trajectory compatible with its behavior can eventually be concatenated with an arbitrary future trajectory. This notion is appealing from many points of view. It does not refer to a particular representation, and, in particular, it applies to systems that are not in state space form. In [12] and [13], mainly finite-dimensional, linear, time-invariant systems have been considered. Also, here we find that in controllable systems any past can be made exactly compatible to any future by a judicious choice of the input over a finite time interval. As may be expected, this property proves to be too demanding for infinite-dimensional systems.

The purpose of this paper is to study controllability using this vantage point for a class of infinite-dimensional systems. Specifically, we study (approximate) controllability for linear systems whose behavior is a shift-invariant, closed, linear subspace

*Received by the editors September 4, 1990; accepted for publication (in revised form) June 14, 1991.

[†]Scuola Normale Superiore, Piazza dei Cavalieri 7, Pisa, 56100, Italy.

[‡]Mathematics Institute, University of Groningen, P.O. Box 800, 9700 AV Groningen, the Netherlands.

of $l^2(\mathbf{Z}, \mathbf{C}^q)$. We also study state representations of such systems and prove a sort of state space isomorphism theorem for almost controllable systems.

The mathematical techniques and methods of proof used here are inspired by functional analytic methods (H^∞ and the like), particularly the work of Fuhrmann [5].

To discuss systems, we follow the so-called behavioral approach, as introduced and developed in [12] and [13]. A *dynamical system* is a triple $\Sigma = (T, W, \mathcal{B})$ with $T \subset \mathbf{R}$ the *time axis*, W the *signal space*, and $\mathcal{B} \subset W^T$ the *behavior*. In this paper, we only consider *discrete-time systems* with $T = \mathbf{Z}$ or *continuous-time systems* with $T = \mathbf{R}$. Moreover, we assume that our systems are *time-invariant*; that is, that $\sigma^t \mathcal{B} = \mathcal{B}$ for every $t \in T$ (*shift-invariance*), where $\sigma^t : W^T \rightarrow W^T$ is the t -*shift* defined by $(\sigma^t f)(t') := f(t + t')$. We also only consider systems with $W = \mathbf{C}^q$ and with \mathcal{B} a linear subspace of W^T (*linear systems*). For most of this paper, we focus on the following class of linear systems:

$$\mathcal{L}_q^2 := \{ \Sigma = (\mathbf{Z}, \mathbf{C}^q, \mathcal{B}) \text{ with } \mathcal{B} \text{ a closed shift-invariant linear subspace of } l_q^2 \},$$

where l_q^2 indicates $l^2(\mathbf{Z}, \mathbf{C}^q)$ the Hilbert space of the \mathbf{C}^q -valued square-summable sequences over \mathbf{Z} . We often refer to a system in \mathcal{L}_q^2 as an l^2 -system.

Example. (1) l^2 -systems defined by input/output maps. Let $T : l_m^2 \rightarrow l_p^2$ be a closed linear map that commutes with the shift σ . T induces the system

$$\Sigma_T := (\mathbf{Z}, \mathbf{C}^{m+p}, G(T)) \in \mathcal{L}_{m+p}^2,$$

where $G(T)$ is the graph of the map T . These input/output systems have been widely investigated in the past (see [5] and [3]); an important case is when T is a convolution operator induced by an l^1 -kernel.

(2) l^2 -systems as restrictions of other systems. To determine how flexible it is to work with systems as a set of trajectories (the behavior), compared with simply input/output relations, suppose that we have linear input/output map $T : (\mathbf{C}^m)^{\mathbf{Z}} \rightarrow (\mathbf{C}^p)^{\mathbf{Z}}$ commuting with the shift. If $T(l_m^2) \not\subset l_p^2$, T does not induce an input/output l^2 -map in the classical sense; nevertheless, we can consider the dynamical system $\Sigma = (\mathbf{Z}, \mathbf{C}^{m+p}, \mathcal{B})$, where $\mathcal{B} := G(T) \cap l_{m+p}^2$. Under certain conditions (for example, when $G(T)$ is closed in the pointwise convergence topology) we have that $\Sigma \in \mathcal{L}_{m+p}^2$ and that Σ completely determines the original behavior $G(T)$. Therefore the theory of l^2 -systems can be used to analyze Σ and thus to infer properties of the map T .

For a given map $w : T \rightarrow W$, we define $w^- := w|_{T \cap (-\infty, 0)}$ (the *past* of w) and $w^+ := w|_{T \cap (0, +\infty)}$ (the *future* of w). If $\mathcal{B} \subset W^T$, we indicate with \mathcal{B}^- and \mathcal{B}^+ the sets of, respectively, the past and the future trajectories of \mathcal{B} .

DEFINITION 1.1. A time-invariant dynamical system $\Sigma = (T, W, \mathcal{B})$ is said to be controllable if, for every w_1 and w_2 in \mathcal{B} , there exist $t' \geq 0$ and $w \in \mathcal{B}$ such that

$$w^- = w_1^- \quad \text{and} \quad (\sigma^{t'} w)^+ = w_2^+.$$

This notion of controllability plays a fundamental role in the theory of linear, time-invariant, finite-dimensional, state space systems, but it proves to be very restrictive when we consider general l^2 -systems, for which we propose the following.

DEFINITION 1.2. A linear time-invariant system $\Sigma = (\mathbf{Z}, \mathbf{C}^q, \mathcal{B}) \in \mathcal{L}_q^2$ is said to be almost controllable if there exists $K > 0$ such that, for every w_1 and w_2 in \mathcal{B} , there exists $v_n \in \mathcal{B}$ for $n = 1, 2, \dots$, yielding the following:

$$(\sigma^{-n} v_n)^- \rightarrow w_1^-, \quad (\sigma^n v_n)^+ \rightarrow w_2^+, \quad \|v_n\|_2 \leq K (\|w_1^-\|_2 + \|w_2^+\|_2),$$

where \rightarrow denotes limit in the l^2 -topology for $n \rightarrow \infty$.

Remark. It is not obvious, from this definition, that a controllable system in \mathcal{L}_q^2 is almost controllable. This is indeed the case and is shown later.

Remark. The uniform boundness requirement on the v'_n s in Definition 1.2 is essential. Indeed, if we drop this, then any system in \mathcal{L}_q^2 would satisfy the property; indeed, let $w_1, w_2 \in \mathcal{B}$ and consider $v_n := \sigma^{-n}w_2 + \sigma^n w_1$. Then $v_n \in \mathcal{B}$ for all n , and it is evident that $(\sigma^n v_n)^+ \rightarrow w_2^+$ and $(\sigma^{-n} v_n)^- \rightarrow w_1^-$ in the l^2 -topology.

2. Controllable systems. The main result of this section concerning l^2 -systems shows that controllable systems in the sense of Definition 1.1 have automatically finite memory.

THEOREM 2.1. *Let $\Sigma = (\mathbf{Z}, \mathbf{C}^q, \mathcal{B})$ be in \mathcal{L}_q^2 . Then Σ is controllable if and only if \mathcal{B} can be expressed as the l^2 -solutions of a linear constant coefficients difference equation; that is, there exists $R(z, z^{-1}) \in \mathbf{C}^{g \times q}[z, z^{-1}]$ such that*

$$(2.1) \quad \mathcal{B} = \{w \in l_q^2 \mid R(\sigma, \sigma^{-1})w = 0\},$$

where $R(\sigma, \sigma^{-1}) : (\mathbf{C}^q)^{\mathbf{Z}} \rightarrow (\mathbf{C}^g)^{\mathbf{Z}}$ is the operator in the shift σ induced by the polynomial matrix $R(z, z^{-1})$.

To prove Theorem 2.1, we must establish a few intermediate results, which have an interest of their own. Also, we work in a somewhat more general setting encompassing l^2 -systems, since we believe that, in this way, a more complete picture of the situation can be drawn without additional effort.

For $w_1, w_2 \in W^T$ and $t \in T$, we denote by the symbol $w_1 \wedge_t w_2$ the *concatenation* of w_1 and w_2 at time t ; i.e., $w_1 \wedge_t w_2(t') := w_1(t')$ for $t' < t$ and $w_1 \wedge_t w_2(t') := w_2(t')$ for $t' \geq t$. We also use the symbol \wedge_t to concatenate restrictions of functions such as, for example, $w_1^- \wedge_0 w_2^+$.

DEFINITION 2.2. Let X be a linear subspace of $(\mathbf{C}^q)^{\mathbf{Z}}$ and let $\|\cdot\|_X$ be a norm on X . $(X, \|\cdot\|_X)$ is said to be a memoryless Banach space if the following hold:

- (1) $(X, \|\cdot\|_X)$ is a complex Banach space,
- (2) X is shift-invariant ($\sigma X = X$) and $\sigma : X \rightarrow X$ is an isometry,
- (3) X is memoryless ($w_1, w_2 \in X \Rightarrow w_1 \wedge_t w_2 \in X$ for all $t \in \mathbf{Z}$).

Remark. If X is a memoryless Banach space, $w \in X$, and $I \subset \mathbf{Z}$, we often identify $w|_I$ with the trajectory in X , which is equal to w on I , and 0 outside of I . Through this identification, the spaces X^- and X^+ are seen as the subspaces of X consisting of the trajectories with support in, respectively, $(-\infty, 0)$ and $[0, +\infty)$. It follows that X^- and X^+ are closed in X and, by condition (3) of the preceding definition, $X = X^- \oplus X^+$. We indicate with P^- and P^+ the linear bounded projections from X on X^- and X^+ , respectively. Once a memoryless Banach space X has been fixed, the convergence of a sequence in the norm of X is simply denoted by the symbol \rightarrow , with no further specification when no confusion can arise.

Example. We now present the following examples of memoryless Banach spaces, which are considered later in the paper:

- (1) The space l_q^p ($1 \leq p < +\infty; q \in \mathbf{N}^+$) of the \mathbf{C}^q -valued sequences over \mathbf{Z} whose p th power is summable, equipped with the norm

$$\|w\|_p := \left(\sum_{-\infty}^{+\infty} |w(t)|_{\mathbf{C}^q}^p \right)^{1/p};$$

(2) The space l_q^∞ ($q \in \mathbf{N}^+$) of the bounded \mathbf{C}^q -valued sequences over \mathbf{Z} , equipped with the norm

$$\|w\|_\infty := \sup_{t \in \mathbf{Z}} |w(t)|_{\mathbf{C}^q};$$

(3) The subspace c_q^0 of l_q^∞ , consisting of the sequences converging to 0 as t approaches $\pm\infty$, equipped with the norm $\|\cdot\|_\infty$.

Note the following chain of inclusions:

$$l_q^1 \subset l_q^p \subset c_q^0 \subset l_q^\infty \quad \forall p \in [1, \infty).$$

If X is a memoryless Banach space contained in $(\mathbf{C}^q)^\mathbf{Z}$, we consider the following class of linear systems:

$$\mathcal{L}_X := \{\Sigma = (\mathbf{Z}, \mathbf{C}^q, \mathcal{B}) \text{ with } \mathcal{B} \text{ a closed shift-invariant linear subspace of } X\}.$$

In the case where $X = l_q^p$, we also use the notation \mathcal{L}_q^p for \mathcal{L}_X .

Let $\Sigma = (T, W, \mathcal{B})$ be a time-invariant system and Δ a positive number. Σ is said to have Δ -finite memory if $w_1, w_2 \in \mathcal{B}$ and $w_1|_{[0, \Delta)} = w_2|_{[0, \Delta)}$ implies that $w_1 \wedge_0 w_2 \in \mathcal{B}$. Σ is said to have finite memory if it has Δ -finite memory for some Δ .

Our first goal is to study the structure of finite memory systems in \mathcal{L}_X . To do this, we must introduce the important system-theoretic concept of completeness. Let $\Sigma = (T, W, \mathcal{B})$ be a time-invariant system; it is said to be complete if, given any $w \in W^T$, we have that $w \in \mathcal{B}$ if and only if $w|_I \in \mathcal{B}|_I$ for every finite interval $I \subset T$ (with obvious meaning of $\mathcal{B}|_I$). The structure of the complete time-invariant linear systems is studied in much detail in [12] and [13]; in particular, there is the following important result.

THEOREM 2.3. *Let $\Sigma = (\mathbf{Z}, \mathbf{C}^q, \mathcal{B})$ be a linear, time-invariant system. The following conditions are then equivalent:*

- (1) Σ is complete,
- (2) $\mathcal{B} \subset (\mathbf{C}^q)^\mathbf{Z}$ is closed in the pointwise convergence topology,
- (3) There exists $R(z, z^{-1}) \in \mathbf{C}^{q \times q}[z, z^{-1}]$ such that $\mathcal{B} = \ker R(\sigma, \sigma^{-1})$.

From (3) of Theorem 2.3, it is clear that any complete linear system over \mathbf{Z} indeed has finite memory. In general, systems in \mathcal{L}_X are not complete; we can actually prove that, if $\mathcal{B} \subset c_q^0$, then Σ is complete if and only if $\mathcal{B} = \{0\}$. Nevertheless, the concept of completeness proves to be useful in our investigation. In fact, we have the following result.

PROPOSITION 2.4. *Let X be a memoryless Banach space contained in c_q^0 and let $\Sigma = (\mathbf{Z}, \mathbf{C}^q, \mathcal{B})$ be in \mathcal{L}_X . Then Σ has finite memory if and only if $\mathcal{B} = \mathcal{B}^{\text{compl}} \cap X$, where $\mathcal{B}^{\text{compl}}$ is the completion of \mathcal{B} (defined as the smallest subspace of $(\mathbf{C}^q)^\mathbf{Z}$ that is shift-invariant, complete, and contains \mathcal{B}).*

Proof. Observe that $\mathcal{B} \subset \mathcal{B}^{\text{compl}} \cap X$. Assume that Σ has Δ -finite memory and let $w \in \mathcal{B}^{\text{compl}} \cap X$. Then there exists a sequence $w_n \in \mathcal{B}$ such that

$$(2.2) \quad w_n|_{[-n, n]} = w|_{[-n, n]} \quad \forall n \in \mathbf{N}.$$

Consider now the linear map $P_\Delta : \mathcal{B}^- \oplus \mathcal{B}^+ \rightarrow \mathcal{B}^-|_{[-\Delta, 0)} \oplus \mathcal{B}^+|_{[0, \Delta)}$ given by

$$P_\Delta(w_1, w_2) = (w_1|_{[-\Delta, 0)}, w_2|_{[0, \Delta)}).$$

Since P_Δ is surjective, there exists a linear map $Q_\Delta : \mathcal{B}^-|_{[-\Delta,0)} \oplus \mathcal{B}^+|_{[0,\Delta)} \rightarrow \mathcal{B}^- \oplus \mathcal{B}^+$ such that $P_\Delta \circ Q_\Delta = Id$. Since Σ has Δ -finite memory, we can assume that

$$\left((\sigma^{-n+\Delta} w_n)^-, (\sigma^{n-\Delta} w_n)^+ \right) = Q_\Delta \left((\sigma^{-n+\Delta} w_n)^-|_{[-\Delta,0)}, (\sigma^{n-\Delta} w_n)^+|_{[0,\Delta)} \right).$$

By (2.2), for n sufficiently large, we then have that

$$(2.3) \quad \left((\sigma^{-n+\Delta} w_n)^-, (\sigma^{n-\Delta} w_n)^+ \right) = Q_\Delta (w|_{[-n,-n+\Delta)}, w|_{[n-\Delta,n)}).$$

Since $w \in X \subset c_q^0$, we have that

$$(2.4) \quad (w|_{[-n,-n+\Delta)}, w|_{[n-\Delta,n)}) \rightarrow 0 \quad \text{as } n \rightarrow +\infty.$$

Q_Δ is bounded, since it acts on a finite-dimensional vector space; therefore, by (2.3) and (2.4), we have that

$$\left((\sigma^{-n+\Delta} w_n)^-, (\sigma^{n-\Delta} w_n)^+ \right) \rightarrow 0 \quad \text{as } n \rightarrow +\infty,$$

which implies, together with (2.2) and condition (2) of Definition 2.2, that $w_n \rightarrow w$. This yields $w \in \mathcal{B}$. The other implication follows from Theorem 2.3. \square

Remark. Proposition 2.4 still holds true if $X = l_q^\infty$ and if we assume that \mathcal{B} is closed in the weak*-topology of l_q^∞ ; the proof is identical.

Let us now state the main result of this section.

THEOREM 2.5. *Let X be a memoryless Banach space and let Σ be a controllable system in \mathcal{L}_X . Then Σ has finite memory.*

We first prove a proposition based on a technical lemma whose proof is omitted since it follows from a straightforward application of the Douglas factorization theorem (see [5]).

LEMMA 2.6. *Let X, Y , and Z be Banach spaces and let $A : X \rightarrow Z$ and $B : Y \rightarrow Z$ be linear bounded maps. If there exists $X_0 \subset X$ subspace of second category in X such that $\mathcal{R}(A|_{X_0}) \subset \mathcal{R}(B)$, then $\mathcal{R}(A) \subset \mathcal{R}(B)$.*

We now state a result that claims that, under certain conditions, controllability may always be achieved in a uniformly bounded finite number of steps, if it can be achieved at all.

PROPOSITION 2.7. *Let X be a memoryless Banach space and let $\Sigma \in \mathcal{L}_X$ be a controllable system. Then there exists $n_0 \in \mathbb{N}$ such that, for all w_1 and w_2 in \mathcal{B} , there exists $w \in \mathcal{B}$ such that*

$$w^- = w_1^- \quad \text{and} \quad (\sigma^{n_0} w)^+ = w_2^+$$

Proof. Let us consider the following sequence of linear bounded maps:

$$(2.5) \quad T_n : \mathcal{B} \rightarrow X^- \oplus X^+,$$

given by $T_n(w) := (w^-, (\sigma^n w)^+)$. By controllability, we have that

$$\bigcup_{n>0} \mathcal{R}(T_n) = \mathcal{B}^- \oplus \mathcal{B}^+.$$

Proposition 2.7 will be proved if we show that there exists $n_0 \in \mathbf{N}$ such that $\mathcal{R}(T_{n_0}) = \mathcal{B}^- \oplus \mathcal{B}^+$. Let us introduce the map

$$T : \mathcal{B} \oplus \mathcal{B} \rightarrow X^- \oplus X^+$$

given by $T(w_1, w_2) := (w_1^-, w_2^+)$. Consider $M_n = T^{-1}\mathcal{R}(T_n)$. Then

$$\bigcup_{n>0} M_n = \mathcal{B} \oplus \mathcal{B},$$

and, since $\mathcal{B} \oplus \mathcal{B}$ is a Banach space, it follows, by a standard category argument (see, for example, [11]) that there exists $n_0 \in \mathbf{N}$ such that M_{n_0} is of second category in $\mathcal{B} \oplus \mathcal{B}$. Applying Lemma 2.6 to the maps T_{n_0} and T , it follows that $\mathcal{R}(T) \subset \mathcal{R}(T_{n_0})$, which implies that $\mathcal{R}(T_{n_0}) = \mathcal{B}^- \oplus \mathcal{B}^+$. \square

More can be said about the range of the map T_{n_0} introduced in (2.5). In fact, consider the map

$$i : \mathcal{B} \rightarrow X^- \oplus X^+ \oplus \mathcal{B}|_{[0, n_0]}$$

with $i = T_{n_0} \oplus P_{n_0}$, where $P_{n_0} : \mathcal{B} \rightarrow \mathcal{B}|_{[0, n_0]}$ is the restriction to the interval $[0, n_0]$. It is clear that i is a linear bounded embedding (injective with closed range) and that P_{n_0} has finite-dimensional range; it is then a standard result from functional analysis (see, for example, [2]) that T_{n_0} also has closed range. This yields the following result.

PROPOSITION 2.8. *Let X be a memoryless Banach space and let $\Sigma = (\mathbf{Z}, \mathbf{C}^q, \mathcal{B}) \in \mathcal{L}_X$ be controllable. Then \mathcal{B}^- and \mathcal{B}^+ are closed subspaces of X .*

Proof of Theorem 2.5. Consider the following subspace of \mathcal{B}^+ :

$$\mathcal{B}_0^+ := \{w^+ \in \mathcal{B}^+ \mid 0 \wedge_0 w^+ \in \mathcal{B}\}.$$

Define the linear map

$$R : \mathcal{B}|_{[0, n_0]} \rightarrow \mathcal{B}^+ / \mathcal{B}_0^+,$$

where n_0 is the same as in Proposition 2.7, by $R(x) = v \pmod{\mathcal{B}_0^+}$, where v is any trajectory in \mathcal{B}^+ such that $0 \wedge_0 x \wedge_{n_0} \sigma^{-n_0} v \in \mathcal{B}$. It is easy to verify that R is a well-defined linear map, and that it is surjective. Since the domain of R is finite-dimensional, it then follows that $\mathcal{B}^+ / \mathcal{B}_0^+$ is also finite-dimensional (this is actually a state space of Σ). Therefore there exists a finite-dimensional subspace N of \mathcal{B}^+ such that $\mathcal{B}^+ = \mathcal{B}_0^+ \oplus N$. Now consider the following decreasing sequence of subspaces of \mathcal{B}^+ :

$$H_n := \{w^+ \in \mathcal{B}^+ \mid w^+|_{[0, n]} = 0\}.$$

Then

$$\bigcap_{n \geq 0} H_n = (0).$$

Consider $K_n := P_N H_n$, where P_N is the projection operator on the subspace N . $\{K_n\}$ is a decreasing sequence of subspaces of N with null intersection; since N has finite dimension, it then follows that there exists $\tilde{n} > 0$ such that $K_{\tilde{n}} = (0)$, which implies that $H_{\tilde{n}} \subset \mathcal{B}_0^+$. We now claim that Σ has \tilde{n} -finite memory; in fact, let $w_1, w_2 \in \mathcal{B}$ such that $w_1|_{[0, \tilde{n}]} = w_2|_{[0, \tilde{n}]}$. Then

$$(w_2 - w_1)^+ \in H_{\tilde{n}} \subset \mathcal{B}_0^+,$$

which implies that

$$w_1 \wedge_0 w_2 = w_1 + (0 \wedge_0 (w_2 - w_1)^+) \in \mathcal{B}. \quad \square$$

We conclude this section with the following summarizing result, which encompasses Theorem 2.1, stated at the beginning of the section.

THEOREM 2.9. *Let X be a memoryless Banach space contained in c_q^0 and let $\Sigma \in \mathcal{L}_X$. Then the following conditions are equivalent:*

- (1) Σ is controllable,
- (2) Σ has finite memory,
- (3) there exists a polynomial matrix $R(z, z^{-1}) \in \mathbb{C}^{g \times q}[z, z^{-1}]$ such that

$$\mathcal{B} = \{w \in X \mid R(\sigma, \sigma^{-1})w = 0\}.$$

Proof. (1) \Rightarrow (2) is Theorem 2.5. (2) \Rightarrow (3) is contained in Proposition 2.4 and Theorem 2.3. Finally, (3) \Rightarrow (1) follows from standard results of the theory of complete systems: [13] and [14] contain a proof for the case where $X = l_q^2$, which is easily generalizable to our case. \square

Remark. The condition that $X \subset c_q^0$ in Theorem 2.9 is essential. In fact, it follows from the results of [13] that Theorem 2.9 is false for l_q^∞ .

3. Almost controllable systems. In this section we specifically consider l^2 -systems, since we believe that the Hilbert structure plays a fundamental role in this context to achieve nice representation results. We make use of frequency domain techniques including Hardy spaces theory; our main references for these matters are [4], [6], and [7].

We start with the following interesting topological characterization of almost controllability.

PROPOSITION 3.1. *Let $\Sigma = (\mathbf{Z}, \mathbb{C}^q, \mathcal{B})$ be an l^2 -system. Then the following two conditions are equivalent:*

- (1) Σ is almost controllable,
- (2) \mathcal{B}^- and \mathcal{B}^+ are closed in l_q^2 .

Proof. (1) \Rightarrow (2). By (1), there exists $k > 0$ such that, for every $w^- \in \mathcal{B}^-$, there exists $v_n \in \mathcal{B}$ such that

$$(\sigma^n v_n)^- \rightarrow w^-, \quad (\sigma^{-n} v_n)^+ \rightarrow 0, \quad \|v_n\|_2 \leq K \|w^-\|_2.$$

Consider $w_n = \sigma^n v_n$; then

$$(3.1) \quad w_n^- \rightarrow w^-,$$

$$(3.2) \quad \|w_n^+\|_2 \leq \|w_n\|_2 \leq K \|w^-\|_2.$$

By (3.2) we can assume, taking a subsequence if necessary, that

$$(3.3) \quad w_n^+ \rightarrow v^+ \in l_q^{2+} \quad \text{weakly.}$$

Equations (3.1) and (3.3) yield

$$w_n = w_n^- \wedge_0 w_n^+ \rightarrow w^- \wedge_0 v^+ \quad \text{weakly,}$$

which implies that

$$(3.4) \quad w^- \wedge_0 v^+ \in \mathcal{B}$$

and, by (3.2),

$$(3.5) \quad \|v^+\| \leq K\|w^-\|_2.$$

From (3.4) and (3.5), it follows, by a standard argument from functional analysis, that $\mathcal{R}(P^-|_{\mathcal{B}}) = \mathcal{B}^-$ is closed. In an analogous way, we see that \mathcal{B}^+ is also closed.

(2) \Rightarrow (1). Since the two projections P^- and P^+ both have closed range, there exists $K > 0$ such that, for all $w^- \in \mathcal{B}^-$ and $w^+ \in \mathcal{B}^+$, there exist w_1 and w_2 in \mathcal{B} such that

$$\begin{aligned} w_1^- &= w^-, & \|w_1\|_2 &\leq K\|w^-\|_2; \\ w_2^+ &= w^+, & \|w_2\|_2 &\leq K\|w^+\|_2. \end{aligned}$$

Now consider $v_n = \sigma^n w_1 + \sigma^{-n} w_2$. Then

$$\begin{aligned} (\sigma^{-n} v_n)^- &= w_1^- + \sigma^{-2n} w_2 \rightarrow w_1^-, \\ (\sigma^n v_n)^+ &= w_2^+ + \sigma^{2n} w_1 \rightarrow w_1^+, \\ \|v_n\|_2 &\leq K(\|w_1^-\| + \|w_2^+\|), \end{aligned}$$

which yields (1). \square

Remark. By Propositions 2.8 and 3.1, it is now evident that, for l^2 -systems, controllability indeed implies almost controllability.

We now study representations of almost controllable systems, and this is the subject of the remainder of this article. We show how it is possible to represent an almost controllable system as the image of l^2 -maps, while, in next section, we study state space representations. The common feature underlying these two representations is the presence of latent variables, namely, variables that are not part of the external signal, but that are introduced to express the internal structure of the system. We return to this point later.

Let $T \subset \mathbf{R}$ and W_1, W_2 be sets; consider $\mathcal{B}_1 \subset W_1^T$ and $\mathcal{B}_2 \subset W_2^T$. A map $F : \mathcal{B}_1 \rightarrow \mathcal{B}_2$ is said to be *causal* if $w_1(t) = w_2(t)$ for all $t \leq t'$ implies that $(Fw_1)(t) = (Fw_2)(t)$ for all $t \leq t'$; F is said to be *anticausal* if $w_1(t) = w_2(t)$ for all $t \geq t'$ implies that $(Fw_1)(t) = (Fw_2)(t)$ for all $t \geq t'$.

The following is the main result of this paper.

THEOREM 3.2. *Let $\Sigma = (\mathbf{Z}, \mathbf{C}^q, \mathcal{B})$ be in \mathcal{L}_q^2 . Then the following conditions are equivalent:*

- (1) Σ is almost controllable,
- (2) \mathcal{B}^- and \mathcal{B}^+ are closed subspaces of l_q^2 ,
- (3) There exist a number $g \in \mathbf{N}$ and two linear bounded maps

$$F^- : l_g^2 \rightarrow l_q^2, \quad F^+ : l_g^2 \rightarrow l_q^2,$$

satisfying the following properties:

- (i) $\mathcal{R}(F^-) = \mathcal{B} = \mathcal{R}(F^+)$,
- (ii) F^- and F^+ commute with σ ,
- (iii) F^- is anticausal and has an anticausal bounded left inverse,
- (iv) F^+ is causal and has a causal bounded left inverse.

Moreover, if any of the three above equivalent conditions is satisfied, then the maps F^- and F^+ in (3) can be chosen to be isometries. If we assume that this is the case, then g is unique, and F^- and F^+ are unique up to right multiplication by unitary isomorphism on \mathbf{C}^q .

The proof of Theorem 3.2 is rather involved. We first discuss some easy aspects.

Proof of Theorem 3.2 (Preamble). Note that the equivalence between (1) and (2) is proved in Proposition 3.1.

Also, it is easy to show that (3) \Rightarrow (2); in fact, consider the maps

$$A : l_g^{2-} \rightarrow l_q^{2-}, \quad A = P^- \circ F^+ \circ P^-$$

and

$$B : l_q^{2-} \rightarrow l_g^{2-}, \quad B = P^- \circ \tilde{F}^+ \circ P^-,$$

where \tilde{F}^+ is the causal left inverse of F^+ , and P^- here indicates both the projection operators on l_g^{2-} and l_q^{2-} . We have that

$$B \circ A = P^- \circ \tilde{F}^+ \circ P^- \circ F^+ \circ P^- = P^- \circ \tilde{F}^+ \circ F^+ \circ P^- = Id|_{l_g^{2-}},$$

which implies that $\mathcal{R}(A) = \mathcal{B}^-$ is closed. In an analogous way, using F^- , it follows that \mathcal{B}^+ is closed.

It therefore remains to be proved that (2) implies (3), and the remainder of this section is devoted to this implication.

Remark. For finite memory l^2 -systems, Theorem 3.2 is already obtained in [14].

We now first introduce the important frequency domain description of an l^2 -system. If $\Sigma = (\mathbf{Z}, \mathbf{C}^q, \mathcal{B}) \in \mathcal{L}_q^2$, consider $\hat{\mathcal{B}}$ the closed subspace of $L_q^2 := L^2(\mathbf{T}, \mathbf{C}^q)$ (the Hilbert space of the \mathbf{C}^q -valued Lebesgue square-integrable functions on the unit circle \mathbf{T}) obtained as the image of \mathcal{B} through the Fourier transform $\mathcal{F}_q : l_q^2 \rightarrow L_q^2$. It is well known that $\hat{\mathcal{B}}$ is a doubly invariant subspace of L_q^2 with respect to the shift $S : L_q^2 \rightarrow L_q^2$, given by $(Sw)(e^{i\theta}) := e^{i\theta}w(e^{i\theta})$. Namely, $S^n\hat{\mathcal{B}} = \hat{\mathcal{B}}$ for all $n \in \mathbf{Z}$. Doubly invariant subspaces of L_q^2 have been widely studied in the past (see [6]); we must recall only a few fundamental facts. A *range function* $J = J(e^{i\theta})$ is a function on the circle \mathbf{T} taking values in \mathcal{G}_q (the family of all the subspaces of \mathbf{C}^q); J is said to be measurable if the orthogonal projection $P(e^{i\theta})$ from \mathbf{C}^q on $J(e^{i\theta})$ is measurable. If J is a measurable range function, we can consider that

$$\mathcal{M}_J = \{\hat{w} \mid \hat{w} \in L_q^2 \text{ and } \hat{w}(e^{i\theta}) \in J(e^{i\theta}) \text{ a.e. on } \mathbf{T}\},$$

and it is easy to show that \mathcal{M}_J is a doubly invariant closed subspace of L_q^2 . A fundamental fact is that all closed, doubly invariant subspaces of L_q^2 are of this form, and also the correspondence between J and \mathcal{M}_J is one-to-one, under the convention that range functions are identified if they are equal almost everywhere. A measurable range function J is called *analytic* if there exists a finite number $\{F_1, \dots, F_g\}$ of elements of H_q^2 (the closed subspace of L_q^2 consisting of the functions whose negative Fourier coefficients are zero) such that $J(e^{i\theta})$ is the span of $\{F_1(e^{i\theta}), \dots, F_g(e^{i\theta})\}$ almost everywhere on \mathbf{T} . In a similar way, using the conjugate space \overline{H}_q^2 , we can introduce the concept of *coanalytic* range function. If J is a range function, we can define the *orthogonal* range function J^\perp by $J^\perp(e^{i\theta}) = (J(e^{i\theta}))^\perp$, where the last orthogonal must be considered in \mathbf{C}^q with respect to the canonical Hermitian inner product; it can be proved that J is analytic if and only if J^\perp is coanalytic.

Let us now introduce the space $L_{g \times q}^\infty$ of the $g \times q$ -matrices of L^∞ -functions defined on \mathbf{T} and the subspace $H_{g \times q}^\infty$ consisting of those whose negative Fourier coefficients are zero. If $F \in L_{g \times q}^\infty$, we will denote by M_F the multiplicative operator induced by F , namely, $M_F : L_q^2 \rightarrow L_g^2$, given by $(M_F w)(e^{i\theta}) := F(e^{i\theta})w(e^{i\theta})$. The following proposition clarifies the relation among all of these concepts. The proof is practically contained in [6]; therefore we only give a sketch of it.

PROPOSITION 3.3. *The following conditions are equivalent:*

- (1) J is an analytic range function,
- (2) There exists $F \in H_{q \times g}^\infty$ such that $\mathcal{M}_J = \mathcal{R}(M_F)$,
- (3) There exists $L \in H_{l \times q}^\infty$ such that $\mathcal{M}_J = \ker(M_L)$.

Moreover, if any of the above equivalent conditions are satisfied, then F in (2) can be chosen to be outer ($\mathcal{M}_J \cap H_q^2 = \mathcal{R}(M_F|_{H_q^2})$) and rigid ($F(e^{i\theta})$ is an isometry almost everywhere). With this choice, g is uniquely determined by the relation $g = \dim J(e^{i\theta})$ almost everywhere, and F is also uniquely determined up to right multiplication by constant unitary matrix.

Proof. (1) \Rightarrow (2) Consider that $\mathcal{A}_J = \mathcal{M}_J \cap H_q^2$. \mathcal{A}_J is a closed S -invariant subspace of H_q^2 ; therefore, by the Beurling–Lax theorem (see [6] and [7]), there exist $g \in \mathbf{N}$ and $F \in H_{q \times g}^\infty$ with F rigid such that $\mathcal{A}_J = FH_g^2$. Since J is an analytic range function, it is evident that $\mathcal{M}_J = \mathcal{R}(M_F)$. Moreover, F is outer by the way it has been defined.

(2) \Rightarrow (1) is trivial.

(3) \Rightarrow (1). Suppose that $\mathcal{M}_J = \ker M_L$. Write L as $L = (L_1, \dots, L_l)^t$, where $L_j \in H_{q \times 1}^\infty$. Then

$$\{w \in \mathcal{M}_J\} \Leftrightarrow \{L_j^t w = 0 \quad \forall j = 1, \dots, l\} \Leftrightarrow \{w \perp \bar{L}_j \quad \forall j = 1, \dots, l\}.$$

Let J' be the coanalytic range function spanned by the family $\{\bar{L}_1, \dots, \bar{L}_l\}$. Since $J = (J')^\perp$, this shows that J is analytic.

Reversing this argument, we see that (1) \Rightarrow (3).

Uniqueness of F and the fact that $g = \dim J(e^{i\theta})$ almost everywhere simply follow from the Beurling–Lax theorem and the fact that F is outer and rigid. \square

Of course, we have the following symmetric result.

PROPOSITION 3.4. *The following conditions are equivalent:*

- (1) J is a coanalytic range function,
- (2) There exists $F \in H_{q \times g}^\infty$ such that $\mathcal{M}_J = \mathcal{R}(M_{\bar{F}})$,
- (3) There exists $L \in H_{l \times q}^\infty$ such that $\mathcal{M}_J = \ker(M_{\bar{L}})$.

Moreover, if any of the above equivalent conditions are satisfied, then F in (2) can be chosen to be outer and rigid. With this choice, g is uniquely determined by the relation $g = \dim J(e^{i\theta})$ almost everywhere, and F is also uniquely determined up to right multiplication by constant unitary matrix.

We are now ready to state and prove the main mathematical result.

LEMMA 3.5. *Let \mathcal{M} be a closed, doubly invariant subspace of L_q^2 . Then the following two conditions are equivalent:*

- (1) \mathcal{M}^- (the projection of \mathcal{M} on $H_q^{2-} := (H_q^2)^\perp$) is closed,
- (2) There exist $F \in H_{q \times g}^\infty$ and $\tilde{F} \in H_{g \times q}^\infty$ such that $\mathcal{M} = \mathcal{R}(M_F)$ and $\tilde{F}F = Id_g$.

Also, if either of these two conditions is satisfied, then F in (2) can be chosen to be rigid and outer.

Proof. (1) \Rightarrow (2). \mathcal{M}^- is closed, and it is invariant for the adjoint of the left shift acting on H_q^{2-} . Therefore, by the Beurling–Lax theorem, there exists a rigid

$\psi \in H_{q \times k}^\infty$, with $k \leq q$, such that

$$(3.6) \quad \mathcal{M}^- = (\bar{\psi} H_k^{2-})^\perp,$$

where the orthogonal is taken with respect to the space H_q^{2-} (and not with respect to all of L_q^2 !).

If $f \in L_q^2$, let us indicate by f^- and f^+ the projections of f on, respectively, H_q^{2-} and H_q^2 . Now, for any $f \in \mathcal{M}$, we have that $f^- \perp \bar{\psi} H_k^{2-}$ by (3.6), and also $f^+ \perp \bar{\psi} H_k^{2-}$. Therefore $f \perp \bar{\psi} H_k^{2-}$ for all $f \in \mathcal{M}$, or, equivalently,

$$(3.7) \quad \psi^t f \perp H_k^{2-} \quad \forall f \in \mathcal{M}.$$

Since \mathcal{M} is a doubly invariant subspace, (3.7) implies that $\psi^t f = 0$ for all $f \in \mathcal{M}$. We now prove that, in fact,

$$(3.8) \quad \psi^t f = 0 \Leftrightarrow f \in \mathcal{M}.$$

Let f be in L_q^2 such that $\psi^t f = 0$; it follows that $\psi^t f^+ + \psi^t f^- = 0$, and therefore $\psi^t f^- \perp H_k^{2-}$, or, equivalently, $f^- \perp \bar{\psi} H_k^{2-}$. By (3.6), it then follows that $f^- \in \mathcal{M}^-$. Since \mathcal{M} is doubly invariant, we also have that

$$\psi^t S^{-n} f = 0 \quad \forall n \in \mathbf{N},$$

which, by the preceding argument, yields

$$(S^{-n} f)^- \in \mathcal{M}^- \quad \forall n \in \mathbf{N}.$$

Therefore there exists a sequence $v_n \in H_q^2$ such that

$$(3.9) \quad (S^{-n} f)^- + v_n \in \mathcal{M} \quad \forall n \in \mathbf{N},$$

and, since \mathcal{M}^- is closed, we can choose v_n such that

$$(3.10) \quad \|v_n\|_2 \leq \|f\|_2,$$

for all n . It follows immediately that $S^n (S^{-n} f)^- \rightarrow f$, and, by (3.10), we can assume—taking, if necessary, a subsequence—that $S^n v_n \rightarrow 0$ weakly. Therefore

$$S^n \left((S^{-n} f)^- + v_n \right) \rightarrow f \quad \text{weakly,}$$

which, by (3.9), implies that $f \in \mathcal{M}$. This yields (3.8), which can be equivalently expressed as $\mathcal{M} = \ker(M_{\psi^t})$. By Proposition 3.3, this implies that there exists $F \in H_{q \times g}^\infty$ such that $\mathcal{M} = \mathcal{R}(M_F)$, and F can be chosen to be outer rigid.

We must still prove that F admits an H^∞ left inverse. This may be seen as follows. Consider the linear bounded map $A : H_g^{2-} \rightarrow H_q^{2-}$, given by

$$A := P^- \circ M_F|_{H_g^{2-}},$$

where P^- denotes the projection onto the subspace H_q^{2-} . Consider the following adjoint of A :

$$A^* : H_q^{2-} \rightarrow H_g^{2-}, \quad A^* = M_{F^*}.$$

Since F is outer and rigid, it is easy to see that A^* is surjective. Consequently, A is injective and has closed range. We now use the fact that a function in a Hardy space $(H_q^2, H_{g \times q}^\infty)$ can be holomorphically extended to the open unit disk D (see [7]); for simplicity of notation, we use the same symbol for a function on \mathbf{T} and its extension to D . If $h \in H_1^\infty$ and $\alpha \in D$, we have that

$$[h(e^{i\theta}) - h(\alpha)](1 - \alpha e^{-i\theta})^{-1} \in H_1^2.$$

Consider that

$$f_\alpha = (1 - |\alpha|^2)^{1/2} (1 - \alpha e^{-i\theta})^{-1}.$$

It is a matter of computation to show that $f_\alpha \in H_1^{2-}$ and $\|f_\alpha\|_2 = 1$ for all $\alpha \in D$. Let $\xi \in \mathbf{C}^g$, with $\|\xi\| = 1$. We can then show that

$$(3.11) \quad A(\xi f_\alpha) = F(\alpha)(\xi f_\alpha).$$

Assume now that there exist sequences $\{\alpha_n\} \subset D$ and $\{\xi_n\} \subset \mathbf{C}^g$, with $\|\xi_n\| = 1$, such that $F(\alpha_n)\xi_n \rightarrow 0$. By (3.11), $A(\xi_n f_{\alpha_n}) \rightarrow 0$, and $\|\xi_n f_{\alpha_n}\|_2 = 1$. This is absurd, since A is injective and has closed range. By the vectorial Corona theorem (see [5]), it then follows that there exists $\tilde{F} \in H_{g \times q}^\infty$ such that $\tilde{F}F = I_g$.

(2) \Rightarrow (1). Simply observe that A admits a left inverse given by $B := P^- \circ M_{\tilde{F}}|_{H_g^{2-}}$.

Therefore $\mathcal{M}^- = \mathcal{R}(A)$ is closed. \square

Naturally, we also have the following symmetric result.

LEMMA 3.6. *Let \mathcal{M} be a closed doubly invariant subspace of L_q^2 . Then the following two conditions are equivalent:*

- (1) \mathcal{M}^+ (the projection of \mathcal{M} on H_q^2) is closed,
- (2) There exist $F \in H_{q \times g}^\infty$ and $\tilde{F} \in H_{g \times q}^\infty$ such that $\mathcal{M} = \mathcal{R}(M_{\tilde{F}})$ and $\tilde{F}F = Id_g$.

Also, if either of these two conditions is satisfied, then F in (2) can be chosen to be rigid and outer.

Proof of Theorem 3.2 (End). (2) \Rightarrow (3). Consider the Fourier transform $\mathcal{F}_q : l_q^2 \rightarrow L_q^2$. If $\mathcal{F}_q(\mathcal{B}) = \mathcal{M}$; then $\mathcal{F}_q(\mathcal{B}^-) = \mathcal{M}^-$, $\mathcal{F}_q(\mathcal{B}^+) = \mathcal{M}^+$, and both are closed in L_q^2 . By Lemmas 3.5 and 3.6, there exist G_1 and G_2 in $H_{q \times g}^\infty$ rigid, outer, both having H^∞ left inverse, and such that

$$\mathcal{R}(M_{G_1}) = \mathcal{M} = \mathcal{R}(M_{\tilde{G}_2})$$

(note that g is the same for G_1 and G_2 by Propositions 3.3 and 3.4). Consider now that

$$F^+ : l_g^2 \rightarrow l_q^2, \quad F^+ := \mathcal{F}_q^{-1} \circ M_{G_1} \circ \mathcal{F}_g$$

and

$$F^- : l_g^2 \rightarrow l_q^2, \quad F^- := \mathcal{F}_q^{-1} \circ M_{\tilde{G}_2} \circ \mathcal{F}_g.$$

Because of standard properties of the Fourier transform, it follows immediately that F^- and F^+ are isometries and that they satisfy properties (i)–(iv) of (3). Finally, the uniqueness of g , F^+ , and F^- also follows from Propositions 3.3 and 3.4. \square

Remark. The representation expressed by condition (3) of Theorem 3.2 is classically known as the *scattering representation*, and it is investigated in [8]. It is worthwhile to note that, while in [8] the scattering representation is derived from the existence of a pair of orthogonal subspaces (the *incoming* and *outgoing* subspaces) of \mathcal{B}

satisfying certain properties, here such a representation is independently derived from the topological assumption expressed by condition (2) of Theorem 3.2.

Remark. It is worthwhile to relate the result of Theorem 3.2 with the result of [13], which states that, if $\Sigma = (\mathbf{Z}, \mathbf{C}^q, \mathcal{B})$ is linear time-invariant and complete, then Σ is controllable if and only if $\mathcal{B} = \mathcal{R}(M(\sigma, \sigma^{-1}))$ for some polynomial matrix $M(s, s^{-1})$. Actually, if this is the case, then there exist polynomial matrices left-invertible $M_1(s)$ and $M_2(s)$ such that $\mathcal{R}(M_1(\sigma)) = \mathcal{B} = \mathcal{R}(M_2(\sigma^{-1}))$. Moreover, $M_1(\sigma)$ and $M_2(\sigma^{-1})$ can be chosen to be injective. In a sense, Theorem 3.2 generalizes this to l^2 -systems.

We conclude this section with an analysis of input/output, almost controllable systems. Let $T : l_m^2 \rightarrow l_p^2$ be a linear bounded map that is causal and commutes with σ ; consider the induced l^2 -system $\Sigma_T = (\mathbf{Z}, \mathbf{C}^{m+p}, G(T))$ (see part (1) of the example in the Introduction). We want to obtain necessary and sufficient conditions on the map T such that Σ_T is almost controllable. Consider now the *Hankel operator* \mathcal{H}_T associated with the map T , namely, $\mathcal{H}_T : l_m^{2-} \rightarrow l_p^2$, given by $\mathcal{H}_T := P^+ \circ T|_{l_m^{2-}}$.

PROPOSITION 3.7. *Σ_T is almost controllable if and only if the Hankel operator \mathcal{H}_T has closed range.*

Proof. It is evident that (t denotes transposition)

$$(3.12) \quad \mathcal{B} = \mathcal{R}([Id_m, T]^t).$$

Note that

$$[Id_m, 0] \circ [Id_m, T]^t = Id_m,$$

which implies, by Lemma 3.5, that \mathcal{B}^- is closed. Therefore by Theorem 3.2, Σ_T is almost controllable if and only if \mathcal{B}^+ is closed. Therefore it suffices to show that \mathcal{B}^+ is closed if and only if \mathcal{H}_T has closed range. Assume that \mathcal{H}_T has closed range and let $f_n \in L_m^2$ be a sequence such that

$$P^+[Id_m, T]^t f_n \rightarrow [\psi_1, \psi_2]^t \in l_m^2 \oplus l_p^2.$$

Then

$$f_n^+ \rightarrow \psi_1, \quad P^+ T f_n \rightarrow \psi_2,$$

which imply that $P^+ T f_n^- \rightarrow \psi_2 - P^+ T \psi_1$. Since \mathcal{H}_T has closed range, it follows that there exists $f^- \in l_m^{2-}$ such that

$$P^+ T f^- = \psi_2 - P^+ T \psi_1,$$

which yields $\psi_2 = P^+ T(\psi_1 + f^-)$. Hence, by (3.12), $[\psi_1, \psi_2]^t \in \mathcal{B}^+$. This shows that \mathcal{B}^+ is closed. On the other hand, if \mathcal{H}_T does not have a closed range, then there exists a sequence $f_n^- \in l_m^{2-}$ such that

$$(3.13) \quad \mathcal{H}_T f_n^- \rightarrow \phi \notin \mathcal{R}(\mathcal{H}_T).$$

There holds that $P^+[Id_m, T]^t f_n^- \rightarrow [0, \phi]^t$. We claim that

$$[0, \phi]^t \notin \mathcal{R}([Id_m, T]^t).$$

Indeed, assume that there exists $f \in l_m^2$ such that

$$P^+[Id_m, T]^t f = [0, \phi]^t$$

This implies that $f^+ = 0$ and $P^+Tf^- = \phi$, which by (3.13) yields a contradiction. This shows that \mathcal{B}^+ is not closed. \square

Remark. In the scalar case ($m=p=1$), the condition for \mathcal{H}_T to have a closed range can be expressed nicely in an equivalent way. Consider that $\hat{T} : L^2 \rightarrow L^2$, given by

$$\hat{T} := \mathcal{F} \circ T \circ \mathcal{F}^{-1}.$$

It is a standard fact that \hat{T} is a multiplicative operator with symbol $H \in H^\infty$ (called the *transfer function* of Σ_T). Note that \mathcal{H}_T is completely determined by H , and it can be proved that \mathcal{H}_T has a closed range if and only if H admits a factorization of the kind $H = \psi \bar{K}$, where $\psi \in H^\infty$ is inner, $K \in H^\infty$, and also there exists $\delta > 0$ such that

$$|\psi(z)| + |K(z)| \geq \delta \quad \forall z \in D.$$

Consequently, a sufficient condition for the almost controllability is, in this case, that H is purely inner or, more generally, that its outer part is rational.

4. Hilbertian state models . We start this section with a few words about general latent variables models, before focusing on state models. A *dynamical system with latent variables* is defined as a quadruple

$$\Sigma_f = (T, W, L, \mathcal{B}_f),$$

with T and W as in the definition of a dynamical system given in the introduction; L is the set of *latent variables*; and $\mathcal{B}_f \subset (W \times L)^T$ the (*full*) *behavior*. As for dynamical systems, we always assume that $T = \mathbf{Z}$ (or $T = \mathbf{R}$) and that our latent variables systems are time-invariant (the definition is analogous to the one for dynamical systems); also we assume linearity, namely, that W and L are vector spaces and \mathcal{B}_f is a linear subspace of $(W \times L)^T$

$$\Sigma = (T, W, P_W \mathcal{B}_f)$$

(where P_W is the projection on the first factor of $W \times L$) is said to be the *manifest* or *external* dynamical system induced by Σ_f ; $P_W \mathcal{B}_f$ is called the *manifest* (or *external*) *behavior*. Σ_f is said to be a *latent variable representation* of Σ . Σ_f is said to be *externally induced* if there exists a map (called the *observability map*)

$$F : P_W \mathcal{B}_f \rightarrow P_L \mathcal{B}_f$$

such that

$$\{(w, g) \in \mathcal{B}_f\} \Leftrightarrow \{w \in P_W \mathcal{B}_f \quad \text{and} \quad g = Fw\}.$$

Σ_f is said to be *past externally induced* (*future externally induced*) if the map F is causal (anticausal).

If $\Sigma = (\mathbf{Z}, \mathbf{C}^q, \mathcal{B})$ is an almost controllable system, the scattering representation of Σ introduced in Theorem 3.2 (3) naturally induces the following two latent variables representations of Σ :

$$\Sigma_f^\pm = \left(\mathbf{Z}, \mathbf{C}^q, \mathbf{C}^g, \mathcal{B}_f^\pm \right),$$

where $\mathcal{B}_f^\pm := \{(w, g) \in l_q^2 \oplus l_g^2 : F^\pm g = w\}$. The existence of a causal (respectively, anticausal) left inverse of F^+ (respectively, F^-) implies that Σ_f^+ (respectively, Σ_f^-) is past (respectively, future) externally induced. Note also that, in both cases, the observability map F (given by the left inverse of, respectively, F^+ and F^- , restricted

to \mathcal{B}) is bounded. Indeed, observe that, whenever Σ and Σ_f are l^2 -systems and Σ_f is an externally induced latent variable representation of Σ , the observability map F has closed graph ($G(F) = \mathcal{B}_f$) and therefore is always bounded.

If $\Sigma = (T, W, \mathcal{B})$ is a dynamical system and $\Sigma_f = (T, W, L, \mathcal{B}_f)$ is a latent variables representation of Σ , then Σ_f is said to be a *state space representation* of Σ if the following holds true:

$$[(w_1, l_1), (w_2, l_2) \in \mathcal{B}_f \text{ and } l_1(t) = l_2(t)] \Rightarrow [(w_1, l_1) \wedge_t (w_2, l_2) \in \mathcal{B}_f].$$

For state space representations, we use the notation Σ_S for Σ_f . In [12] and [13], a general theory of state space representations of a dynamical system is developed, and a notion of complexity is introduced, as well as a notion of equivalence. In particular, it is proved that, if Σ is a linear system, then the linear time-invariant state space representations of Σ , of minimal complexity, are all equivalent to each other; moreover, it is shown how to canonically construct a minimal state space representation. However, when we study dynamical systems carrying a topological structure on the behavior (as l^2 -systems), then it is of interest to consider topological structures on the state space, also, and, consequently, to have notions of complexity and equivalence where these topological concepts are also considered. One of the main effects of this new setting is the loss of the equivalence of all the state space representations of minimal complexity, even for a linear system. The main result of this section is to show that, for almost controllable l^2 -systems, this equivalence is actually preserved! We start with an interesting definition, which induces a topological structure on state space representations.

DEFINITION 4.1. Let $\Sigma = (\mathbf{Z}, \mathbf{C}^q, \mathcal{B}) \in \mathcal{L}_q^2$ and let $\Sigma_S = (\mathbf{Z}, \mathbf{C}^q, X, \mathcal{B}_S)$ be a time-invariant state space representation of Σ , with X a complex separable Hilbert space. Σ_S is said to be a Hilbertian state space representation of Σ if the following condition holds true: For every A open subset of l_q^2 such that $\mathcal{B} \subset A$, there exists an open neighborhood N of 0 in X such that

$$(4.1) \quad [(w_1, x_1), (w_2, x_2) \in \mathcal{B}_S \text{ and } x_1(0) - x_2(0) \in N] \Rightarrow [w_1 \wedge_0 w_2 \in A].$$

Condition (4.1) simply says that if two trajectories in \mathcal{B} have states that at $t = 0$ are “very close” to each other, then the concatenation of these two trajectories will also be “very close” to \mathcal{B} .

We denote by H_Σ the set of all the Hilbertian state space representations of the l^2 -system Σ . If $\Sigma_S = (\mathbf{Z}, \mathbf{C}^q, X, \mathcal{B}_S) \in H_\Sigma$, define

$$(4.2) \quad X^{\text{eff}} := \{\xi \in X \mid \exists (w, x) \in \mathcal{B}_S \text{ such that } x(0) = \xi\}.$$

X^{eff} is a subspace of X (not necessarily closed), and it is called the *effective state space* of Σ_S .

DEFINITION 4.2. $\Sigma_S \in H_\Sigma$ is said to be trim if $X^{\text{eff}} = X$; it is said to be almost trim if $\overline{X^{\text{eff}}} = X$.

If $\Sigma_S \in H_\Sigma$ is externally induced, then we can define a linear map

$$(4.3) \quad g : \mathcal{B} \rightarrow X,$$

given by $g(w) := x(0)$, where $x \in X^{\mathbf{Z}}$ is such that $(w, x) \in \mathcal{B}_S$. With a slight abuse of notation, we also call g the observability map of Σ_S .

DEFINITION 4.3. $\Sigma_S \in H_\Sigma$ is said to be boundedly externally induced if g is bounded.

We now give some examples of such state space representations.

Example. (1) If $\Sigma = (\mathbf{Z}, \mathbf{C}^q, \mathcal{B}) \in \mathcal{L}_q^2$, consider that

$$(4.4) \quad \Sigma_S^{\text{trivial}} := (\mathbf{Z}, \mathbf{C}^q, \mathcal{B}, \mathcal{B}_S^{\text{trivial}}),$$

where

$$\mathcal{B}_S^{\text{trivial}} := \{(w, x) | w \in \mathcal{B} \text{ and } x(t) = \sigma^t w\}.$$

$\Sigma_S^{\text{trivial}}$ is usually called the *trivial state space representation* of Σ , and it is immediate to see that $\Sigma_S^{\text{trivial}} \in H_\Sigma$, is trim, and boundedly externally induced.

(2) A more important state space representation of Σ is the following:

$$(4.5) \quad \Sigma_S^c := (\mathbf{Z}, \mathbf{C}^q, \mathcal{B}/_D, \mathcal{B}_S^c),$$

where

$$D := \{w \in \mathcal{B} \mid w \wedge_0 0 \in \mathcal{B}\}$$

and

$$\mathcal{B}_S^c := \{(w, x) \mid w \in \mathcal{B} \text{ and } x(t) = \sigma^t w \pmod{D}\}.$$

It is called the *canonical state space representation* of Σ . It is well known [13] that Σ_S^c is a trim, past and future externally induced state space representation of Σ . It is easy to see that Σ_S^c is also boundedly externally induced. Moreover, $\Sigma_S^c \in H_\Sigma$ if we consider $\mathcal{B}/_D$ with the natural quotient structure, after noting that D is closed in \mathcal{B} . Indeed, fix (w_1, x_1) and (w_2, x_2) in \mathcal{B}_S^c , and assume that

$$\|x_1(0) - x_2(0)\| \leq \delta.$$

This means that there exists $v \in D$ such that $\|w_1 - w_2 + v\|_{\mathcal{B}} \leq \delta$, or, also, that

$$\|(w_1 + 0 \wedge_0 v) - (w_2 - v \wedge_0 0)\|_{\mathcal{B}} \leq \delta.$$

Now $w_1 \wedge_0 w_2 = (w_1 + 0 \wedge_0 v) - (w_2 - v \wedge_0 0)$, and, therefore,

$$\|w_1 \wedge_0 w_2 - w_1 + 0 \wedge_0 v\|_{\mathcal{B}} \leq \delta.$$

This shows that $\Sigma_S \in H_\Sigma$.

DEFINITION 4.4. Let $\Sigma_S^i = (\mathbf{Z}, \mathbf{C}^q, X_i, \mathcal{B}_S^i)$ be in H_Σ for $i = 1, 2$. Σ_S^1 is said to be more complex than Σ_S^2 ($\Sigma_S^1 \geq \Sigma_S^2$) if there exists a linear bounded surjective map $f : X_1 \rightarrow X_2$ such that, for every $(w, x_2) \in \mathcal{B}_S^2$, there exists $x_1 \in X_1^{\mathbf{Z}}$ such that $(w, x_1) \in \mathcal{B}_S^1$ and $f \circ x_1 = x_2$.

DEFINITION 4.5. Let Σ_S^1 and Σ_S^2 as in Definition 4.4. Σ_S^1 is said to be equivalent to Σ_S^2 ($\Sigma_S^1 \simeq \Sigma_S^2$) if there exists a linear bounded bijective map $f : X_1 \rightarrow X_2$ such that $(w, x_1) \in \mathcal{B}_S^1$ if and only if $(w, f \circ x_1) \in \mathcal{B}_S^2$.

Note that \geq is a preorder on H_Σ , while \simeq is an equivalence relation.

We indicate with H_Σ^* the set of all the minimal elements of H_Σ with respect to the pre-order \geq ; namely, $\Sigma_S \in H_\Sigma^*$ if and only if $\Sigma_S \in H_\Sigma$ and $[\Sigma_S \geq \Sigma'_S] \Rightarrow [\Sigma_S \simeq \Sigma'_S]$. We later show that the canonical representation Σ_S^c is always minimal and that, if Σ is almost controllable, then any other minimal representation is, in fact, equivalent to Σ_S^c .

Let us now investigate in more detail the continuity requirement in Definition 4.1 of a Hilbertian state space representation. Assume that $\Sigma_S = (\mathbf{Z}, \mathbf{C}^q, X, \mathcal{B}_S)$ is a linear time-invariant state space representation of $\Sigma = (\mathbf{Z}, \mathbf{C}^q, \mathcal{B})$, and assume that X is a complex separable Hilbert space. Define the map

$$(4.6) \quad \psi(\Sigma_S) : X^{\text{eff}} \rightarrow l_q^2 / \mathcal{B}$$

by

$$\psi(\Sigma_S)(x) := w_1 \wedge_0 w_2 \pmod{\mathcal{B}},$$

where w_1, w_2 is any pair of trajectories in \mathcal{B} such that there exist x_1 and x_2 in $X^{\mathbf{Z}}$ with $(w_i, x_i) \in \mathcal{B}_S$ for $i = 1, 2$ and $x_1(0) - x_2(0) = x$. To better understand how the map $\psi(\Sigma_S)$ really acts on X^{eff} , observe that the codomain of $\psi(\Sigma_S)$ can be canonically identified with $l_q^{2+} / \mathcal{B}_0^+$, where

$$\mathcal{B}_0^+ := \{w^+ \in l_q^{2+} \text{ such that } 0 \wedge_0 w^+ \in \mathcal{B}\}.$$

Through this identification, $\psi(\Sigma_S)$ acts as follows: Given $x \in X^{\text{eff}}$, $\psi(\Sigma_S)(x)$ is the equivalence class $(\text{mod } \mathcal{B}_0^+)$ of all the possible futures of the system \mathcal{B} compatible with initial state at time $t = 0$ equal to x . An analogous identification can be made with respect to the past.

LEMMA 4.6. $\psi(\Sigma_S)$ is a well-defined linear map.

Proof. Let (w'_i, x'_i) and (w''_i, x''_i) be in \mathcal{B}_S for $i = 1, 2$ and assume that

$$(4.7) \quad x'_1(0) - x'_2(0) = x = x''_1(0) - x''_2(0).$$

Consider $(w'_i - w''_i, x'_i - x''_i)$ for $i = 1, 2$ and observe that, by (4.7),

$$(x'_1(0) - x''_1(0)) - (x'_2(0) - x''_2(0)) = 0.$$

Therefore $(w'_1 - w''_1) \wedge_0 (w'_2 - w''_2) \in \mathcal{B}$, or, equivalently,

$$w'_1 \wedge_0 w'_2 - w''_1 \wedge_0 w''_2 \in \mathcal{B},$$

which shows that $\psi(\Sigma_S)$ is well defined. A straightforward calculation shows that $\psi(\Sigma_S)$ is linear. \square

Using this lemma we can obtain the following nice characterization of Hilbertian state space models.

PROPOSITION 4.7. Σ_S is in H_Σ if and only if $\psi(\Sigma_S)$ is bounded.

Proof. The proof is an immediate application of the definition of H_Σ . \square

If $\Sigma_S \in H_\Sigma$, then $\psi(\Sigma_S)$, being bounded on X^{eff} , can be extended in a unique way to a linear bounded map acting on $\overline{X^{\text{eff}}}$; for simplicity of notation, we denote this extension also by the symbol $\psi(\Sigma_S)$.

If $\Sigma_S = (\mathbf{Z}, \mathbf{C}^q, X, \mathcal{B}_S) \in H_\Sigma$, denote $\psi := \psi(\Sigma_S)$ and consider that

$$(4.8) \quad \Sigma'_S := (\mathbf{Z}, \mathbf{C}^q, X / \ker \psi, \mathcal{B}'_S),$$

where

$$\mathcal{B}'_S := \left\{ (w, \bar{x}) \in \left(\mathbf{C}^q \oplus X / \ker \psi \right)^{\mathbf{Z}} \mid (w, x) \in \mathcal{B}_S \right\},$$

where \bar{x} denotes the equivalence class of $x \pmod{\ker \psi}$. We have the following result.

PROPOSITION 4.8. Σ'_S is a Hilbertian externally induced state space representation. Moreover, $\Sigma_S \geq \Sigma'_S$.

Proof. Let us prove that Σ'_S is a state space representation of Σ . Fix (w_1, x_1) and (w_2, x_2) in \mathcal{B}_S , and assume that

$$x_0 := x_1(0) - x_2(0) \in \ker \psi.$$

We must only prove that

$$(4.9) \quad (w_1, \bar{x}_1) \wedge_0 (w_2, \bar{x}_2) \in \mathcal{B}'_S.$$

Since $x_0 \in \ker \psi \cap X^{\text{eff}}$, there exists $(w, x) \in \mathcal{B}_S$ such that $x(0) = x_0$ and $w \wedge_0 0 \in \mathcal{B}$. Let $x' \in X^{\mathbf{Z}}$ be such that

$$(4.10) \quad (w \wedge_0 0, x') \in \mathcal{B}_S$$

and let $x'' = x - x'$; then

$$(4.11) \quad (0 \wedge_0 w, x'') \in \mathcal{B}_S.$$

Consider now

$$(w_1 + (0 \wedge_0 w), x_1 + x'') \quad \text{and} \quad (w_2 - (w \wedge_0 0), x_2 - x').$$

These are elements of \mathcal{B}_S and $(x_1 + x'')(0) - (x_2 - x')(0) = 0$. Therefore

$$(4.12) \quad \begin{aligned} & (w_1, x_1 + x'') \bigwedge_0 (w_2, x_2 - x') \\ &= (w_1 + (0 \wedge_0 w), x_1 + x'') \bigwedge_0 (w_2 - (w \wedge_0 0), x_2 - x') \in \mathcal{B}_S. \end{aligned}$$

By (4.10) and (4.11), it is evident that $\psi(x'(t)) = 0$ for all $t \geq 0$ and $\psi(x''(t)) = 0$ for all $t \leq 0$; this, together with (4.12), yields (4.9).

The fact that $\Sigma'_S \in H_\Sigma$ follows from the commutativity of the following diagram:

$$\begin{array}{ccc} X^{\text{eff}} & & \xrightarrow{\psi} l_q^2/\mathcal{B} \\ \downarrow \pi & & \nearrow \psi' \\ X^{\text{eff}}/\ker \psi \cap X^{\text{eff}} & & \end{array}$$

where $\psi' := \psi(\Sigma'_S)$.

To prove that Σ'_S is externally induced, assume that $(0, \bar{x}) \in \mathcal{B}'_S$; then $x(t) \in \ker \psi$ for all $t \in \mathbf{Z}$, which implies that $\bar{x} = 0$.

Finally, the projection

$$\pi : X \rightarrow X/\ker \psi$$

yields $\Sigma_S \geq \Sigma'_S$ in the sense of Definition 4.4. \square

PROPOSITION 4.9. Let $\Sigma_S = (\mathbf{Z}, \mathbf{C}^q, X, \mathcal{B}_S) \in H_\Sigma$. The following conditions are then equivalent:

- (1) $\Sigma_S \in H_\Sigma^*$,
- (2) Σ_S is almost trim and $\psi(\Sigma_S)$ is injective on X .

Proof. (1) \Rightarrow (2). Consider that

$$\Sigma_S^{\text{eff}} := \left(\mathbf{Z}, \mathbf{C}^q, \overline{X^{\text{eff}}}, \mathcal{B}_S \right).$$

It is evident that $\Sigma_S \geq \Sigma_S^{\text{eff}}$, which yields $\Sigma_S \simeq \Sigma_S^{\text{eff}}$. This shows that Σ_S is almost trim, since Σ_S^{eff} is. Analogously, by Proposition 4.8, $\Sigma_S \simeq \Sigma'_S$, where Σ'_S has been defined in (4.8). Therefore there exists an isomorphism

$$f : X \rightarrow X / \ker \psi(\Sigma_S)$$

such that

$$\psi(\Sigma'_S) \circ f = \psi(\Sigma_S),$$

which proves that $\psi(\Sigma_S)$ is injective.

(2) \Rightarrow (1). Assume that there exists $\tilde{\Sigma}_S = \left(\mathbf{Z}, \mathbf{C}^q, \tilde{X}, \tilde{\mathcal{B}}_S \right) \in H_\Sigma$ such that $\Sigma_S \geq \tilde{\Sigma}_S$. From Proposition 4.8, it follows that Σ_S and $\tilde{\Sigma}_S$ are externally induced; moreover, we have that the following diagram commutes:

$$(4.13) \quad \begin{array}{ccc} \mathcal{B} & \xrightarrow{g} & X \\ \downarrow \tilde{g} & \swarrow f & \\ \tilde{X} & & \end{array}$$

where g (respectively, \tilde{g}) are the observability maps of Σ_S (respectively, $\tilde{\Sigma}_S$) as defined in (4.3), and f is the linear bounded surjective map yielding the preorder \geq between Σ_S and $\tilde{\Sigma}_S$. Fix now $(w, x) \in \mathcal{B}_S$; by (4.13) it follows that $(w, f \circ x) \in \tilde{\mathcal{B}}_S$. It is then clear, by Definition 4.5, that to prove that $\Sigma_S \simeq \tilde{\Sigma}_S$, it suffices to prove that the map f is injective. To prove this, consider the following diagram:

$$\begin{array}{ccccc} \mathcal{B} & \xrightarrow{g} & X & \xrightarrow{\psi} & l_q^2 / \mathcal{B} \\ & \searrow \tilde{g} & \downarrow f & \nearrow \tilde{\psi} & \\ & & \tilde{X} & & \end{array}$$

where $\psi := \psi(\Sigma_S)$ and $\tilde{\psi} := \psi(\tilde{\Sigma}_S)$. It is evident that $\psi \circ g = \tilde{\psi} \circ \tilde{g}$. Using the commutativity of (4.13), we obtain that $\psi \circ g = \tilde{\psi} \circ f \circ g$, which implies that

$$\psi(x) = \left(\tilde{\psi} \circ f \right)(x) \quad \forall x \in X^{\text{eff}}.$$

Since $\overline{X^{\text{eff}}} = X$ and since all the maps involved are bounded, it follows that $\psi = \tilde{\psi} \circ f$. Since ψ is injective, this shows the injectivity of f , as desired. \square

COROLLARY 4.10. *It holds that $\Sigma_S^c \in H_\Sigma^*$.*

Proof. Σ_S^c is trim; therefore, by Proposition 4.9, we must only prove that $\psi^c := \psi(\Sigma_S^c)$ is injective on X . Observe that

$$\psi^c : \mathcal{B}/D \rightarrow l_q^2 / \mathcal{B}$$

is given by $\psi^c(w \pmod{D}) = (w \wedge_0 0) \pmod{\mathcal{B}}$. Therefore

$$\psi^c(w \pmod{D}) = 0 \iff w \wedge_0 0 \in \mathcal{B} \iff$$

$$\Longleftrightarrow w \in D \Longleftrightarrow w \pmod{D} = 0. \quad \square$$

PROPOSITION 4.11. *Let Σ_S be in H_Σ . The following conditions are then equivalent:*

- (1) Σ is minimal, trim, and boundedly externally induced,
- (2) $\Sigma_S \simeq \Sigma_S^c$.

Proof. (1) \Rightarrow (2). Let us denote by X the state space of Σ_S , and by A that of Σ_S^c . For $x \in X$, consider that

$$\mathcal{B}(x) := \{w \in \mathcal{B} \mid \exists z \in X^{\mathbf{Z}} : (w, z) \in \mathcal{B}_S \text{ and } z(0) = x\}.$$

Similarly, define $\mathcal{B}(a)$ for $a \in A$. It is easy to see, since \mathcal{B}_S^c is past and future externally induced, that, for every $x \in X$, there exists one and only one $a \in A$ such that $\mathcal{B}(x) \subset \mathcal{B}(a)$. This yields the existence of a linear surjective map $f : X \rightarrow A$ such that $(w, z) \in \mathcal{B}_S$ if and only if $(w, f \circ z) \in \mathcal{B}_S^c$. We now prove that f is bounded. Consider the following commutative diagram:

$$(4.14) \quad \begin{array}{ccc} \mathcal{B} & \xrightarrow{g} & X \\ \downarrow g_c \swarrow f & & \\ A & & \end{array},$$

where g (respectively, g_c) are the observability maps of Σ_S (respectively, Σ_S^c). Let $C \subset A$ be an open set. We have that

$$f^{-1}(C) = g(g_c^{-1}(C)).$$

Since g_c is bounded and g is open (it is surjective and bounded by (1)), it follows that $f^{-1}(C)$ is open in X . Therefore f is bounded and $\Sigma_S \geq \Sigma_S^c$. Since Σ_S is minimal, it follows that $\Sigma_S \simeq \Sigma_S^c$.

(2) \Rightarrow (1) is contained in Corollary 4.10. \square

We now focus on almost controllable systems.

PROPOSITION 4.12. *Let $\Sigma = (\mathbf{Z}, \mathbf{C}^q, \mathcal{B}) \in \mathcal{L}_q^2$ be almost controllable. Then any minimal Hilbertian state space representation Σ_S of Σ is trim and boundedly externally induced.*

Proof. Consider that $\psi := \psi(\Sigma_S)$, as defined before. It is evident that

$$\mathcal{R}(\psi|_{X^{\text{eff}}}) = \mathcal{B}^- \wedge_0 \mathcal{B}^+ / \mathcal{B},$$

which is, by the assumption of almost controllability, closed in l_q^2 / \mathcal{B} . Since $\overline{X^{\text{eff}}} = X$, it follows that

$$\mathcal{R}(\psi) = \mathcal{B}^- \wedge_0 \mathcal{B}^+ / \mathcal{B}.$$

Since ψ is injective, this implies that $X^{\text{eff}} = X$.

By Proposition 4.8, Σ_S is externally induced. We then have the following commutative diagram:

$$\begin{array}{ccc} \mathcal{B} & \xrightarrow{g} & X \\ \downarrow \phi \swarrow \psi & & \\ \mathcal{B}^- \wedge_0 \mathcal{B}^+ / \mathcal{B} & & \end{array},$$

where $\phi(w) := w \wedge_0 0 \pmod{\mathcal{B}}$. Since ϕ is bounded and ψ is an isomorphism, it follows that g is bounded. This completes the proof. \square

We now state the main result of this section. It consists of a state space isomorphism theorem for Hilbert space systems, with the state space isomorphism induced by bounded linear maps. Note that almost controllability plays an essential role in this result!

THEOREM 4.13. *Let $\Sigma \in \mathcal{L}_q^2$ be almost controllable and let $\Sigma_S \in H_\Sigma$. Then the following conditions are equivalent:*

- (1) $\Sigma_S \in H_\Sigma^*$,
- (2) $\Sigma_S \simeq \Sigma_S^c$,
- (3) Σ_S is trim and past and future externally induced.

Proof. (1) \Rightarrow (2) follows from Propositions 4.11 and 4.12. (2) \Rightarrow (3) follows from Proposition 4.11 and the definition of Σ_S^c . Finally, (3) \Rightarrow (1) follows from Proposition 4.9 and the evident fact that, if Σ_S satisfies (3), then $\psi(\Sigma_S)$ is injective. \square

As already mentioned, almost controllability is essential to have the isomorphism result expressed in Theorem 4.9. In fact, we have the following proposition.

PROPOSITION 4.14. *Let $\Sigma = (\mathbf{Z}, \mathbf{C}^q, \mathcal{B}) \in \mathcal{L}_q^2$ be not almost controllable. Then there exists $\Sigma_S \in H_\Sigma^*$, which is not trim.*

Proof. Consider the following state space representation of Σ :

$$\Sigma_S = (\mathbf{Z}, \mathbf{C}^q, X, \mathcal{B}_S),$$

where

$$X := \overline{\mathcal{B}^- \wedge_0 \mathcal{B}^+} / \mathcal{B}$$

and

$$\mathcal{B}_S := \left\{ (w, x) \in (\mathbf{C}^q \oplus X)^{\mathbf{Z}} \mid w \in \mathcal{B} \text{ and } x(t) = (\sigma^t w) \wedge_0 0 \pmod{\mathcal{B}} \right\}.$$

It is easy to check that this is indeed a state space representation of Σ and $\psi(\Sigma_S)$ is simply the inclusion map on l_q^2/\mathcal{B} . It then follows that $\Sigma_S \in H_\Sigma^*$; on the other hand, Σ_S is not trim, since either \mathcal{B}^- or \mathcal{B}^+ is not closed. \square

Remark. If $\Sigma \in \mathcal{L}_q^2$ is almost controllable, then the minimal state space representation Σ_S^c can be represented in the following familiar way. There exist

$$A : X \rightarrow X, \quad B : \mathbf{C}^g \rightarrow X, \quad C : X \rightarrow \mathbf{C}^q, \quad D : \mathbf{C}^g \rightarrow \mathbf{C}^q$$

linear bounded maps yielding the following representation: $(w, x) \in \mathcal{B}_S^c$ if and only if there exists $v \in l_g^2$ such that

$$\sigma x = Ax + Bv, \quad w = Cx + Dv.$$

Such a representation is called a *driving variable representation*. The details of the construction of such a representation are not presented here, since it is completely analogous to the so-called shift realization that has been investigated for input/output systems in [5].

We close our study of state space representations by a discussion of the relation between our concepts of controllability and the classical concept of state controllability. In [12] we have defined state point controllability as the possibility of transferring the system between any two states in finite time. The appropriate version of almost state point controllability proves to be the following.

DEFINITION 4.15. Let $\Sigma \in \mathcal{L}_q^2$ and $\Sigma_S = (\mathbf{Z}, \mathbf{C}^q, X, \mathcal{B}_S) \in H_\Sigma$. Σ_S is said to be almost state point controllable if there exists $K > 0$ such that, for every pair of elements x_1 and x_2 in X , there exists a sequence $(v_n, y_n) \in \mathcal{B}_S$ yielding the following:

$$y_n(-n) \rightarrow x_1, \quad y_n(n) \rightarrow x_2, \quad \|v_n\|_2 \leq K (\|x_1\|^- + \|x_2\|^+),$$

where the convergence is in the Hilbertian topology of the space X , and where $\|\cdot\|^-$ and $\|\cdot\|^+$ are defined as follows:

$$(4.15) \quad \begin{aligned} \|x\|^- &:= \inf \{ \|v^-\|_2 \mid \exists y \in X^{\mathbf{Z}} \text{ with } (v, y) \in \mathcal{B}_S \text{ and } y(0) = x \}, \\ \|x\|^+ &:= \inf \{ \|v^+\|_2 \mid \exists y \in X^{\mathbf{Z}} \text{ with } (v, y) \in \mathcal{B}_S \text{ and } y(0) = x \}. \end{aligned}$$

It is possible to prove that, for minimal state space representations, almost controllability and almost state point controllability are indeed equivalent.

PROPOSITION 4.16. Let $\Sigma = (\mathbf{Z}, \mathbf{C}^q, \mathcal{B}) \in \mathcal{L}_q^2$ and let $\Sigma_S = (\mathbf{Z}, \mathbf{C}^q, X, \mathcal{B}_S) \in H_\Sigma^*$. Then the following conditions are equivalent:

- (1) Σ is almost controllable,
- (2) Σ_S is almost state point controllable.

Proof. (1) \Rightarrow (2). By Theorem 4.13 we can assume, without loss of generality, that $\Sigma_S = \Sigma_S^c$. Let x_1 and x_2 be in X . Then there exist w_1 and w_2 in \mathcal{B} such that

$$x_i = w_i \pmod{D} \quad \text{for } i = 1, 2$$

and

$$(4.16) \quad \|w_1^-\| \leq 2\|x_1\|^- , \quad \|w_2^+\| \leq 2\|x_2\|^+ .$$

By (1), there exists a sequence $v_n \in \mathcal{B}$ such that

$$(4.17) \quad (\sigma^{-n}v_n)^- \rightarrow w_1^-, \quad (\sigma^n v_n)^+ \rightarrow w_2^+$$

and

$$(4.18) \quad \|v_n\|_2 \leq K (\|w_1^-\| + \|w_2^+\|),$$

where K is a positive constant depending only on Σ . Now, consider $y_n \in X^{\mathbf{Z}}$, given by

$$y_n(t) = \sigma^t v_n \pmod{D}.$$

By (4.17) and by the fact that \mathcal{B}^- is closed (see Proposition 3.1), it follows that $y_n(-n) \rightarrow x_1$. Analogously, $y_n(n) \rightarrow x_2$. By (4.16) and (4.18),

$$\|v_n\|_2 \leq 2K (\|x_1\|^- + \|x_2\|^+).$$

This yields (2).

(2) \Rightarrow (1). Let w_1 and w_2 be in \mathcal{B} and let x_1 and $x_2 \in X^{\mathbf{Z}}$ be such that $(w_i, x_i) \in \mathcal{B}_S$ for $i = 1, 2$. By (2) there exists a sequence $(y_n, v_n) \in \mathcal{B}_S$ such that

$$(4.19) \quad y_n(n) \rightarrow x_2(0), \quad y_n(-n) \rightarrow x_1(0)$$

and

$$(4.20) \quad \|v_n\|_2 \leq K (\|x_1(0)\|^- + \|x_2(0)\|^+).$$

Consider that

$$(4.21) \quad z_n = \sigma^n w_1 \wedge_{-n} v_n \wedge_n \sigma^{-n} w_2 \in l_q^2.$$

By (4.19), and by the fact that $\Sigma_S \in H_\Sigma$, there exists $\tilde{z}_n \in \mathcal{B}$ such that

$$\|\tilde{z}_n - z_n\| \rightarrow 0 \quad \text{for } n \rightarrow +\infty.$$

In particular, $(\sigma^{-n} \tilde{z}_n)^- - (\sigma^{-n} z_n)^- \rightarrow 0$, which implies, by (4.20) and (4.21), that

$$(4.22) \quad (\sigma^{-n} \tilde{z}_n)^- \rightarrow w_1^-.$$

In a similar way, we can prove that

$$(4.23) \quad (\sigma^n \tilde{z}_n)^+ \rightarrow w_2^+.$$

By (4.20) and (4.21), we also have that

$$\|z_n\| \leq (1 + K) (\|w_1^-\| + \|w_2^+\|),$$

and, on the other hand, it is not restrictive to assume that $\|\tilde{z}_n\| \leq 2\|z_n\|$. This, together with (4.22) and (4.23), yields (1). \square

Classically, of course, controllability is always studied for systems with inputs. We now briefly analyze the concept of almost state point controllability for state space representations of causal input/output l^2 -systems, and we establish a relation with the classical notion of exact controllability as considered, for example, in [4] and [5].

Let $T : l_m^2 \rightarrow l_p^2$ be a linear bounded causal map commuting with the shift and let

$$\Sigma_T = (\mathbf{Z}, \mathbf{C}^{m+p}, G(T))$$

be the induced l^2 -system as defined in part (1) of the example in the Introduction. Let Σ_S be in H_{Σ_T} and assume that it is past externally induced. It is then possible to consider the following linear map (the *reachability map* of Σ_S):

$$(4.24) \quad R : (l_m^2)^- \rightarrow X,$$

given by

$$Rv^- := g \begin{pmatrix} v^- \\ Tv^- \end{pmatrix},$$

where g is the observability map defined in (4.3). As in [5], we call a state space representation *exactly state point controllable* if R is bounded and surjective. If Σ_S is trim and almost state point controllable, then Σ_S is exactly state point controllable. In fact, in this case, Σ_T is almost controllable by Proposition 4.16, and an easy argument using the commutative diagram (4.14) shows that Σ_S is boundedly externally induced. This yields that Σ_S is exactly state point controllable. In particular, by Theorem 4.13, it follows that, if Σ_S is minimal and almost state point controllable, then Σ_S is exactly state point controllable.

On the other hand, exact state point controllability does not, in general, imply almost state point controllability; it is easy, in fact, to see that the canonical representation Σ_S^c is always exactly state point controllable, but, by Proposition 4.16, it is almost state point controllable if and only if Σ_T is almost controllable. Nevertheless,

with the additional assumption that the norm $\|\cdot\|^+$ (see (4.15)) is equivalent to the original norm $\|\cdot\|_X$ of X as a Hilbert space, then exact state point controllability implies almost state point controllability. In fact, using the facts that R is an open map and that the two norms are equivalent, it is easy to prove that, for every $x \in X$, there exists $(v, y) \in \mathcal{B}_S$ such that

$$(4.25) \quad y(0) = x, \quad y(-n) \rightarrow 0, \quad \|v\|_2 \leq K\|y\|^+$$

for a suitable constant $K > 0$. On the other hand \mathcal{B}_- is closed by Proposition 3.1, and this implies that, for every $x \in X$, there exists a sequence $(v_n, y_n) \in \mathcal{B}_S$ such that

$$(4.26) \quad y_n(0) \rightarrow x, \quad y_n(n) \rightarrow 0, \quad \|v_n\|_2 \leq K'\|x\|_-,$$

where K' is a suitable positive constant. It is evident that (4.25) and (4.26) yield almost state point controllability. Let us conclude by noting that the two norms $\|\cdot\|^+$ and $\|\cdot\|_X$ are indeed equivalent for the so-called "restricted shift" state space representations, which have been investigated in [5].

5. Conclusions and extensions. In this paper we have investigated the notion of controllability as the possibility of concatenation of arbitrary trajectories. For discrete-time systems, we have seen that the possibility of concatenation of trajectories in finite time requires the system to have finite memory, which is equivalent to it having a finite-dimensional state space representation. For infinite-dimensional systems, therefore, we introduced the notion of almost controllability. Our main result is Theorem 3.2, where it is shown that almost controllability is equivalent to the existence of a scattering representation.

As a first application of almost controllability, we obtained in Theorem 4.13 a state space isomorphism result for almost controllable systems. Also, we related our notion of controllability to the classical notion of state point controllability. Under suitable conditions, these notions indeed prove to be equivalent.

Many of the results presented here for the discrete-time case are actually extendable to the continuous time case: in particular, §3 on the representation of almost controllable systems, and §4 on state models. On the other hand, the characterization of the controllable systems in the continuous-time case is more involved and still incomplete. In particular, it is reasonable to conjecture that finite memory and controllability will also be equivalent here. However, in this case, finite memory is not equivalent to the existence of a finite-dimensional state representation.

We believe there are two extensions worth investigating: (i) constructing a representation theory in the fashion of §3 and §4 for l^2 -systems where autonomous phenomena are present, and investigating systems embedded in other memoryless Banach structures: of particular interest it would be to work with behaviors \mathcal{B} in l_q^∞ .

REFERENCES

- [1] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.
- [2] R. E. EDWARDS, *Functional Analysis*, Holt, Rinehart, and Winston, New York, 1965.
- [3] Y. FOURES AND I. E. SEGAL, *Causality and Analyticity*, Trans. Amer. Math. Soc., 78 (1955), pp. 385–405.
- [4] P. FUHRMANN, *Exact controllability and observability and realization theory*, J. Math. Anal. Appl., 53 (1976), pp. 385–405.
- [5] ———, *Linear Systems and Operators in Hilbert Space*, McGraw-Hill, New York, 1981.
- [6] H. HELSON, *Lectures on Invariant Subspaces*, Academic Press, New York, 1964.

- [7] K. HOFFMAN, *Banach Spaces of Analytic Functions*, Prentice-Hall, Englewood Cliffs, NJ, 1962.
- [8] P. D. LAX AND R. S. PHILLIPS, *Scattering Theory*, Academic Press, New York, 1967.
- [9] J. L. LIONS, *Optimal Control of Systems Described by Partial Differential Equations*, Springer, New York, 1971.
- [10] H. H. ROSENBROCK, *State Space and Multivariable Theory*, John Wiley, New York, 1970.
- [11] W. RUDIN, *Functional Analysis*, McGraw-Hill, New York, 1973.
- [12] J. C. WILLEMS, *From time series to linear systems*, Part I: *Finite dimensional linear time-invariant systems*, *Automatica*, 22 (1986), pp. 561–580.
- [13] ———, *Models for dynamics*, *Dynamics Reported*, Vol. 2 (1989).
- [14] J. C. WILLEMS AND C. HEIJ, *l (sub 2) -systems and their scattering representation*, in *Operator Theory and Systems*, H. Bart, I. Gohberg, and M. A. Kaashoek, eds., Birkhäuser Verlag, Berlin, New York, 1986, pp. 443–448.

ALGEBRAIC DIFFERENTIAL EQUATIONS AND RATIONAL CONTROL SYSTEMS*

YUAN WANG[†] AND EDUARDO D. SONTAG[‡]

Abstract. An equivalence is shown between realizability of input/output (i/o) operators by rational control systems and high-order algebraic differential equations for i/o pairs. This generalizes, to nonlinear systems, the equivalence between autoregressive representations and finite-dimensional linear realizability.

Key words. rational systems, input/output equations, identification

AMS(MOS) subject classifications. 93B15, 93A25, 93B25, 93B27, 93B29

1. Introduction. In this paper we prove an equivalence between realizability of input/output (i/o) operators by rational control systems and the existence of high-order algebraic differential equations relating derivatives of inputs and outputs.

In many experimental situations involving systems, it is often the case that one can model system behavior through differential equations, which are referred to as i/o equations in this work, of the type

$$(1) \quad E\left(u(t), u'(t), u''(t), \dots, u^{(r)}(t), y(t), y'(t), y''(t), \dots, y^{(r)}(t)\right) = 0,$$

where $u(\cdot)$ and $y(\cdot)$ are the input and output signals, respectively, and E is a polynomial. An i/o operator $F: u(\cdot) \mapsto y(\cdot)$ is said to *satisfy* (1) if the equation holds for each sufficiently differentiable input u and the corresponding output $y = F[u]$ of F . (Precise definitions are given later.)

The functional relation E is usually estimated, for instance, through least squares techniques, if a parametric general form (e.g., polynomials of fixed degree) is chosen. For example, in linear systems theory, we often deal with degree-one polynomials E , below:

$$(2) \quad y^{(k)}(t) = a_1 y(t) + \dots + a_k y^{(k-1)}(t) + b_1 u(t) + \dots + b_k u^{(k-1)}(t)$$

(or their frequency-domain equivalent, transfer functions; the difference equation analogue is sometimes called an “autoregressive moving average” representation). In the linear case, such representations form the basis of much of modern systems analysis and identification theory.

State-space formalisms are more popular than i/o equations in nonlinear control, however. There, we assume that inputs and outputs are related by a system of *first*-order differential equations

$$(3) \quad x'(t) = f(x(t)) + G(x(t))u(t), \quad y(t) = h(x(t)),$$

where the state $x(t)$ is now a vector, and no derivatives of controls are allowed. These descriptions are central to the modern nonlinear control theory, as they permit the application of techniques from differential equations, dynamical systems, and optimization theory. Thus a basic question is that of deciding when a given i/o

* Received by the editors August 20, 1990; accepted for publication (in revised form) April 12, 1991. This research was supported in part by United States Air Force grant AFOSR-88-0235.

[†] Mathematics Department, Florida Atlantic University, Boca Raton, Florida 33431.

[‡] Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903.

operator admits a representation of this form. This is the area of *realization theory*, which is closely related, especially when stochastic effects are included, to *systems identification*. Roughly speaking, if such a state space description does exist for a given i/o operator, then we say that the i/o operator is *realizable*. More precisely, we are interested in realizations in which the entries of f and G , as well as the function h , can be expressed in terms of rational functions of the state, but, due to the technical problems that arise in the definition because of possible poles of these rational functions, we give the precise definition in terms of “singular polynomial systems,” and we also study realizability by (nonsingular) polynomial systems.

We know that an equation such as (2) can be reduced by adding state variables for enough derivatives of the output y to a system (3) of first-order equations, with $f(x)$ linear and $G(x)$ constant, i.e., a linear finite-dimensional system. In frequency-domain terms, rationality of the transfer function is equivalent to realizability. (For references on the linear theory, see, e.g., [14], [23], and [32].) One of the methods for obtaining a linear realization from a given linear i/o equation relies on Lord Kelvin’s principle for solving differential equations by means of mechanical analogue computers (cf. [14]). The principle, which was suggested 100 years ago, provided a way for simulating a system without using differentiators.

For nonlinear systems, this reduction presents a far harder problem, one that is, to a great extent, unsolved. The problem is basically that of, in some sense, replacing a nontrivial equation (1) by a system of first-order equations (3), which does not involve derivatives of the inputs. A number of results were already available about the relation between (1) and (3); see, for instance, [4], [12], or [26]. It is easy to show, by elementary arguments involving finite transcendence degree, that any i/o operator realizable by a rational state space system satisfies some i/o equation of type (1), with E a polynomial. In [6] it was remarked—as a consequence of theorems from differential algebra—that to characterize the i/o behavior of a state space system *uniquely*, we must add inequality constraints to (1). In [18] and [27] it was shown that, under some constant rank conditions, the outputs of an observable smooth state space system can be described by an equation of type (1) for which E is a smooth function, and local i/o equations were shown to exist, for generic initial states of (3) in [3].

1.1. Our approach. The discrete-time work reported in [20] and [21] provided one approach to relating these two types of representations—with difference equations appearing instead—based on the idea of dealing with existence of realizations separately from the question of “wellposedness” of the equation (in the sense to be described). This work has been developed further, and it was, for example, used as a basis of identification algorithms by other authors; see, for instance, [15] and [5]. (The former reference shows also how to include stochastic effects.) These results have recently been extended to continuous-time for the very special case of *bilinear* systems: A theorem showed that realizability by such systems is equivalent to the existence of an E of a special form, namely, affine on y (see [22]). However, the techniques in [22] were linear-algebraic and hence not powerful enough to handle the extension of [21] to the general nonlinear case. The present work completes the development of the extension of the main realizability result in [21] to continuous-time.

The separation into “wellposedness” and realizability can be illustrated with the simple example $u(t)y'(t) = 1$. This can never be satisfied by all the i/o pairs corresponding to a state space system, as remarked in [22]. Moreover, it cannot even be satisfied by any “input/output map” of the type that we consider, realizable or

not. Indeed, our main result shows that if the equation is well posed, in the sense that it is an equation satisfied by all i/o pairs corresponding to what we call a *Fliess operator*—i.e., one described by a convergent generating series—and if E is a polynomial, then it is always realizable by a singular polynomial system, or a rational system with possible poles. (Singular systems appear naturally in control theory, for instance, in robotics; see [17] for many examples.) In the special case when (1) is recursive—i.e., the coefficient of the highest derivative of y in (1) does not depend on the lower derivatives of y —our construction provides a polynomial realization (no poles).

Our formalism is based on the *generating series* suggested by Fliess in the late 1970s, who was, in turn, motivated by Chen's work on power series solutions of differential equations. The i/o operators induced by convergent generating series form a very general class of causal operators, capable of representing a variety of nonlinear systems. We call them "Fliess operators." For instance, any i/o operator induced by an initialized analytic state space system affine in controls can be described in this manner. In [29], we develop the basic analytic properties of Fliess operators, and results from there are freely used here.

The proofs are based on a careful analysis of the concept of *observation space*, introduced in [16] (and [21] for discrete-time), developed further in [11], and later rediscovered by many authors. One of the central technical results relates two different definitions of this space: one in terms of smooth controls, and another in terms of piecewise constant ones. These two definitions are seen to coincide. One of them immediately relates to i/o equations, while the other is related to realizability through the notion of *observation algebras* and *observation fields*. The latter are the analogues of the corresponding discrete-time concepts studied in [21]. For differential equations, they were first employed in [1] and [2]; the results there related finiteness properties of the various algebraic objects to realizability, in strict analogy to the relations that hold in discrete time [21].

In addition to single operators, it is also natural to study *families of i/o maps*, defined by a *family of convergent generating series*. To study a single i/o map is natural as a formal description of an initialized *black box*, but, in general, a system may induce more than one i/o map. For example, a system described by an ordinary differential equation on a manifold may induce infinitely many i/o maps, each of them corresponding to some initial state. We should study all the i/o maps induced by the system simultaneously rather than individually, unless a fixed initial state is of particular interest. This leads to the concept of families of i/o maps. One question arises naturally: When can a family of i/o maps be realized by *one* state space system; i.e., when can all the members of the family be realized by some singular polynomial system in such a way that each member of the family is associated to some initial state of the system? We prove that a family of i/o maps is realizable in this sense if and only if all the members of the family satisfy a common i/o equation.

The paper is organized as follows. After introducing an algebraic structure on series, the shuffle product, we consider observation spaces. Then we study i/o equations satisfied by i/o operators, showing that the existence of an i/o equation implies that the observation field is a finitely generated field extension of \mathbb{R} . In the next section, realizability by polynomial systems and singular polynomial systems is considered; the result there is that realizability by singular polynomial systems is guaranteed by the condition that the observation field is a finitely generated extension of \mathbb{R} . The approach pursued there is to use the generators of the field as state variables and use

the equalities that hold among the generators to construct the needed vector fields. In the main section, based on the previous results, we establish the equivalence between equations and realizability. We also show there that a special kind of equations, recursive i/o equations, lead to realization by polynomial systems. However, as opposed to the general case, the converse of this fact is not true in general, and a counterexample is provided to illustrate the fact that realizability by a polynomial system may not lead to a recursive i/o equation. Finally, we extend our main result to families of i/o operators.

This paper is heavily algebraic. All analytic properties needed are quoted from [28] and [29] and are not proved here. The latter paper also shows how, using analytic function theory, as well as differential-geometric nonlinear realization tools, an analogous theory can be developed for local realizability provided that an equation with E analytic (not necessarily polynomial) exist for the given operator.

2. Preliminaries. Let m be a fixed integer and consider the “alphabet” set

$$P = \{\eta_0, \eta_1, \dots, \eta_m\}$$

and P^* , the free monoid generated by P , where the neutral element of P^* is the empty word, denoted by 1, and the product is concatenation. Let

$$P^k = \{\eta_{i_1}\eta_{i_2}\cdots\eta_{i_k} : 0 \leq i_s \leq m, 1 \leq s \leq k\}$$

for each $k \geq 0$. We define \mathcal{P} to be the \mathbb{R} -algebra generated by P^* , i.e., the set of all polynomials in the variables η_i 's. A power series in the noncommutative variables $\eta_0, \eta_1, \dots, \eta_m$ is a formal power series

$$(4) \quad c = \sum_{\iota \in I^*} \langle c, \eta_\iota \rangle \eta_\iota,$$

where $\eta_\iota = \eta_{i_1}\eta_{i_2}\cdots\eta_{i_\iota}$, if $\iota = i_1i_2\cdots i_\iota$, and $\langle c, \eta_\iota \rangle \in \mathbb{R}$ for each multi-index ι . Note that c is a polynomial if and only if there are only finitely many $\langle c, \eta_\iota \rangle$'s that are nonzero. A power series is nothing more than a mapping from I^* to \mathbb{R} ; as we see later, however, the algebraic structures suggested by the series formalism are very important. We use \mathcal{S} to denote the set of all power series (over a fixed but arbitrary alphabet P).

For $c, d \in \mathcal{S}$ and $\gamma \in \mathbb{R}$, $\gamma c + d$ is the series defined as follows:

$$\langle \gamma c + d, \eta_\iota \rangle = \gamma \langle c, \eta_\iota \rangle + \langle d, \eta_\iota \rangle.$$

With these operations, \mathcal{S} forms a vector space over \mathbb{R} . In addition, we can introduce an algebra structure on \mathcal{S} by defining the *shuffle product* on \mathcal{S} . First, we define the shuffle product on words

$$\sqcup : P^* \times P^* \longrightarrow P^*$$

inductively on length in the following way:

$$1 \sqcup \eta = \eta \sqcup 1 = \eta \quad \text{for any } \eta \in P,$$

$$(5) \quad \eta_i \eta_\iota \sqcup \eta_j \eta_\kappa = \eta_i (\eta_\iota \sqcup \eta_j \eta_\kappa) + \eta_j (\eta_i \eta_\iota \sqcup \eta_\kappa) \quad \text{for any } \eta_i, \eta_\kappa \in P^*, \eta_j, \eta_\iota \in P.$$

It can be proved by induction that an equivalent way to define the shuffle product is to replace (5) by the following:

$$(6) \quad \eta_i \eta_j \sqcup \eta_\kappa \eta_j = (\eta_i \sqcup \eta_\kappa \eta_j) \eta_j + (\eta_i \eta_j \sqcup \eta_\kappa) \eta_j \quad \text{for any } \eta_i, \eta_\kappa \in P^*, \eta_j \in P.$$

Then we extend the shuffle product to power series in the following way. For

$$c = \sum \langle c, \eta_i \rangle \eta_i \quad \text{and} \quad d = \sum \langle d, \eta_\kappa \rangle \eta_\kappa,$$

we define

$$(7) \quad c \sqcup d = \sum \langle c, \eta_i \rangle \langle d, \eta_\kappa \rangle \eta_i \sqcup \eta_\kappa.$$

With the operations “+” and “ \sqcup ” defined as above, \mathcal{S} forms a commutative \mathbb{R} -algebra.

Remark 2.1. We can also define a comultiplication $M : \mathcal{S} \rightarrow \mathcal{S} \times \mathcal{S}$ and a counit ε over \mathcal{S} . First, for $z \in P^*$, define

$$M(z) = \sum_{z_1 z_2 = z} (z_1, z_2),$$

$$\varepsilon(z) = \begin{cases} 0 & \text{if } z \neq 1, \\ 1 & \text{if } z = 1. \end{cases}$$

Then extend M and ε to \mathcal{S} . It can be shown that \mathcal{S} forms a Hopf algebra with the antipode σ defined by

$$\sigma(\eta_{i_1} \eta_{i_2} \cdots \eta_{i_s}) = (-1)^s \eta_{i_s} \cdots \eta_{i_2} \eta_{i_1}$$

for any s and $\eta_{i_1} \eta_{i_2} \cdots \eta_{i_s} \in P^*$ (cf. [25]). Though \mathcal{S} possesses both an algebra structure and a coalgebra structure, in this work, however, only the algebra structure of \mathcal{S} is studied. \square

LEMMA 2.2. *The algebra \mathcal{S} is an integral domain.*

Proof. First, we order the basis elements $(\eta_{i_1}, \dots, \eta_{i_k})$ of P^* lexicographically with respect to k, i_1, i_2, \dots, i_k . Then take two nonzero series c and d and let

$$z_1 = \eta_{i_1} \cdots \eta_{i_m} \quad \text{and} \quad z_2 = \eta_{j_1} \cdots \eta_{j_n}$$

be the smallest basis element of P^* appearing in c and d , respectively, with nonzero coefficients. Let $w := \eta_{l_1} \cdots \eta_{l_{m+n}}$ be the smallest basis elements of P^* appearing in $z_1 \sqcup z_2$. Then the coefficient of w in $c \sqcup d$ is

$$\langle c \sqcup d, w \rangle = \sum_{\iota, \kappa} \langle c, \eta_\iota \rangle \langle d, \eta_\kappa \rangle \langle \eta_\iota \sqcup \eta_\kappa, w \rangle.$$

Using the minimality property of w, z_1, z_2 , we obtain that

$$\langle c \sqcup d, w \rangle = \langle c, z_1 \rangle \langle d, z_2 \rangle \langle z_1 \sqcup z_2, w \rangle,$$

which is nonzero, since $\langle c, z_1 \rangle, \langle d, z_2 \rangle, \langle z_1 \sqcup z_2, w \rangle$ are all nonzero. \square

The method used in the above proof is similar to the method used in [19], where the author proved that the ring of polynomials in $\eta_0, \eta_1, \dots, \eta_m$ is an integral domain.

In [19] the author used the greatest basis elements (the “degree”) for polynomials, while here we use the smallest basis elements (the “order”) for power series. Alternatively, we could prove this elementary fact by establishing an isomorphism with a ring of power series in (infinitely many) *commuting* variables, along the lines of the discussion in pp. 46–47 in [21].

To define operators associated to series, we need a notion of convergence. We follow [8], [13], and [29] and say that c is *convergent* if there exist some nonnegative real numbers K and M so that the estimate

$$(8) \quad |\langle c, \eta_\iota \rangle| \leq KM^k k!$$

holds for each multi-index $\iota \in I^k$ and each $k \geq 0$. As in [29], we denote by \mathcal{U}_T the set of all essentially bounded measurable functions $u : [0, T] \rightarrow \mathbb{R}^m$, for each fixed $T > 0$. It is convenient to think of \mathcal{U}_T as a space with the L_1 norm ($\|u\|_1 := \max\{\|u_i\|_1 : 1 \leq i \leq m\}$), but we also, at times, use the norm $u_\infty := \max\{\|u_i\|_\infty : 1 \leq i \leq m\}$.

By induction of l , we define, for each input $u \in \mathcal{U}_T$, and each $\iota \in I^l$,

$$(9) \quad V_\phi := 1, \quad V_{i_1 \dots i_{l+1}}[u](t) = \int_0^t u_{i_1}(s) V_{i_2 \dots i_{l+1}}(s) ds.$$

Here u_i denotes the i th coordinate of u , if $i = 1, 2, \dots, m$, and we make the convention $u_0(t) \equiv 1$. Using these notations, to each convergent power series c in $\eta_0, \eta_1, \dots, \eta_m$, we can associate the i/o *operator*

$$(10) \quad F_c[u](t) = \sum \langle c, \eta_\iota \rangle V_\iota[u](t).$$

This is well defined for any T admissible for c , i.e. $T < (Mm + M)^{-1}$; see for [8], [13], and [29] for details (series (10) converges uniformly and absolutely for all $t \in [0, T]$ and all those $u \in \mathcal{U}_T$ such that $\|u\|_\infty < 1$; we denote $\mathcal{V}_T = \{u \in \mathcal{U}_T : \|u\|_\infty < 1\}$, the set of all such controls).

The correspondence between series and operators is one-to-one in the following sense. Assume that c and d are two convergent series, and F_c coincides with F_d on \mathcal{V}_T for some $T > 0$; then the two power series c and d coincide. See [30], [29] for these facts as well as further properties of generating series and their associated operators.

Assume that c and d are two convergent power series and T is admissible for both c and d ; then T is admissible for both $c + d$ and $c \sqcup d$ (cf. [28]). Now for any positive integer n , denote

$$c^n = \underbrace{c \sqcup c \sqcup \dots \sqcup c}_n,$$

and $c^0 = 1$. In [7] it was shown that, for any polynomial $p \in \mathbb{R}[X_1, X_2, \dots, X_s]$ and any s convergent power series c_1, \dots, c_s ,

$$(11) \quad p(F_{c_1}, F_{c_2}, \dots, F_{c_s}) = F_{p(c_1, c_2, \dots, c_s)};$$

that is, the assignment $c \mapsto F_c$ is a homomorphism from the set of all convergent series, seen as an algebra under the shuffle product, into the set of i/o operators (more precisely, identifying operators with their restrictions to smaller time intervals). By the previous discussion, this homomorphism is one-to-one.

Assume that c is a convergent series and pick up a T admissible for c . We show in [29] that F_c is a continuous operator from \mathcal{V}_T to $C[0, T]$ with respect to the L_1 norm

in \mathcal{V}_T and the \mathcal{C}^0 norm in $\mathcal{C}[0, T]$. Furthermore, F_c maps functions of class \mathcal{C}^{k-1} to functions of class \mathcal{C}^k , for all $k = 1, 2, \dots$, and analytic functions to analytic functions. See also [10] for the proof of the following formula:

$$(12) \quad \frac{d}{dt} F_c[u](t) = F_{\eta_0^{-1}c}[u](t) + \sum_{j=1}^m u_j(t) F_{\eta_j^{-1}c}[u](t),$$

where $\langle z^{-1}c, \eta_l \rangle := \langle c, z\eta_l \rangle$ is defined for each $z \in P^*$ and each $\eta_l \in P^*$. (It is known, cf. [22], that $z^{-1}c$ is convergent if c is, and, in fact, the same T remains admissible.)

3. Observation space. In realization theory and many other areas of nonlinear control, the concept of observation space plays a central role. Observation spaces were first defined in [16] and [11] for continuous-time systems and, in [21], for discrete-time. The solution of many problems for systems, such as the “bilinear immersion” problem treated in [11], are characterized by properties of these spaces. We may define observation spaces in two very different ways, as discussed in this section. Roughly, one possibility is to take the functions corresponding to derivatives with respect to switching times in piecewise constant controls, and the other is to take high-order derivatives at the final time, if smooth controls are used. We show, however, that both definitions lead to the same concept, and this equivalence provides one of the main technical tools that we use to establish the main result.

For each power series c , we define the first type of observation space \mathcal{F}_1 as the linear subspace of the set of all power series spanned by all the elements of the form $z^{-1}c$, i.e.,

$$(13) \quad \mathcal{F}_1(c) = \text{span}_{\mathbb{R}}\{z^{-1}c : z \in P^*\}.$$

Then $F_1(c)$ consists of convergent series if c is a convergent series (cf. [22]).

For a convergent power series c , the elements of $\mathcal{F}_1(c)$ are closely related to the derivatives of $F_c[u]$ with respect to switching times in piecewise constant controls, in the sense to be made precise next.

For any $\mu \in \mathbb{R}^m$, we define $P^\mu : \mathbf{F} \rightarrow \mathbf{F}$, where \mathbf{F} is the set of all germs of i/o operators induced by convergent generating series, in the following way:

$$(P^\mu \circ F_c)[u](t) = \left. \frac{d}{dt} \right|_{\tau=0^+} F_c[u \#_t \omega^\mu](t + \tau),$$

where $u \#_t v$ denotes the concatenated control

$$(u \#_t v)(\sigma) = \begin{cases} u(\sigma) & \text{if } 0 \leq \sigma \leq t, \\ v(\sigma - t) & \text{if } t < \sigma \leq T \end{cases}$$

for any u and v , and $\omega^\mu(\tau) \equiv \mu$, a constant control. Note that $(P^\mu \circ F_c)[u]$ is defined if u is in the domain of F_c . In fact, by formula (12), we have the following easy relation:

$$P^\mu \circ F_c = F_{\eta_0^{-1}c} + \sum_{j=1}^m \mu_j F_{\eta_j^{-1}c},$$

for any $\mu = (\mu_1, \mu_2, \dots, \mu_m) \in \mathbb{R}^m$.

For a convergent power series c , let $\mathcal{G}_1(c)$ be the smallest subspace of operators that contains F_c and that is invariant under P^μ for any $\mu \in \mathbb{R}^m$. By Lemma 2.1 in [30], $\mathcal{G}_1(c)$ is isomorphic to $\mathcal{F}_1(c)$.

To introduce the second type of observation space, we must introduce more notations. Consider, for each $q \geq 1$, the following set of $2 \times q$ matrices:

$$(14) \quad S_q = \left\{ \begin{pmatrix} j_1 & j_2 & \cdots & j_q \\ i_1 & i_2 & \cdots & i_q \end{pmatrix} : \right. \\ \left. i_s, j_s \in \mathbb{Z}, 1 \leq i_s \leq m, j \geq 0, (1, 0) \leq (i_1, j_1) \leq \cdots \leq (i_q, j_q) \right\},$$

where “ \leq ” is the lexicographic order on the set $\{(i, j) : i, j \in \mathbb{Z}\}$. For each element

$$\begin{pmatrix} j_1 & j_2 & \cdots & j_q \\ i_1 & i_2 & \cdots & i_q \end{pmatrix}$$

in S_q and each $n \geq q + \sum j_r$, we define

$$(15) \quad \Gamma_{i_1 \dots i_q}^{j_1 \dots j_q}(n) = \eta_0^{(k)} \sqcup \eta_{i_1} X^{(j_1)} \sqcup \eta_{i_2} X^{(j_2)} \sqcup \cdots \sqcup \eta_{i_q} X^{(j_q)} \Big|_{X=1},$$

where $k = n - q - \sum j_s$. The evaluation is interpreted as follows. First, introduce a new variable X , then perform all shuffles, and finally delete X from the result. Note that (15) is different from $\eta_{i_1} \sqcup \eta_{i_2} \sqcup \cdots \sqcup \eta_{i_q}$; for example,

$$\eta_0 \sqcup \eta_1 X \Big|_{X=1} = \eta_0 \eta_1 + 2\eta_1 \eta_0,$$

while

$$\eta_0 \sqcup \eta_1 = \eta_0 \eta_1 + \eta_1 \eta_0.$$

For any word $w \in P^*$ and each series $c \in \mathcal{S}$, we define $\psi_c(w) = w^{-1}c$, and, more generally, for any polynomial $d = \sum \langle d, \eta_\kappa \rangle \eta_\kappa$, we let

$$\psi_c(d) = \sum \langle d, \eta_\kappa \rangle \eta_\kappa^{-1} c.$$

Now let $X_j = (X_{1j}, \dots, X_{mj})$ be m indeterminates over \mathbb{R} , for $j \geq 0$. For any $n > 0$, let

$$(16) \quad c_n(X_0, \dots, X_{n-1}) = \psi_c(\eta_0^{(n)}) + \sum_{q=1}^n \sum \frac{1}{s_1! \cdots s_p!} \psi_c \left(\Gamma_{i_1 \dots i_q}^{j_1 \dots j_q}(n) \right) X_{i_1 j_1} \cdots X_{i_q j_q},$$

where the second sum is taken over the set of all those

$$\begin{pmatrix} j_1 & j_2 & \cdots & j_q \\ i_1 & i_2 & \cdots & i_q \end{pmatrix} \in S_q$$

such that $\sum j_s + q \leq n$, and where s_1, \dots, s_p are integers, so that

$$\begin{pmatrix} j_1 & j_2 & \cdots & j_q \\ i_1 & i_2 & \cdots & i_q \end{pmatrix} = \left(\underbrace{\alpha_1 \cdots \alpha_1}_{s_1} \quad \underbrace{\alpha_2 \cdots \alpha_2}_{s_2} \quad \cdots \quad \underbrace{\alpha_p \cdots \alpha_p}_{s_p} \right)$$

and $(\alpha_1, \beta_1) < (\alpha_2, \beta_2) < \cdots < (\alpha_p, \beta_p)$. For $n = 0$, we simply define $c_0 := c$. It was shown in [30] that, for each integer n and every $u \in \mathcal{V}_T$ such that T is admissible for c , we have that

$$(17) \quad \frac{d^n}{dt^n} F_c[u](t) = F_{c_n(u(t), \dots, u^{n-1}(t))}[u](t).$$

Hence, for any $\mu_0, \dots, \mu_{n-1} \in \mathbb{R}^m$,

$$(18) \quad \left. \frac{d^n}{d\tau^n} \right|_{\tau=0+} F_c[u \#_t w_\mu](t + \tau) = F_{c_n(\mu_0, \dots, \mu_{n-1})}[u](t),$$

where $w_\mu(t) = \mu_0 + \mu_1 t + \dots + \mu_{s-1} (t^{s-1}/(s-1)!)$.

The second type of observation space associated to c , $\mathcal{F}_2(c)$, is defined as follows:

$$(19) \quad \mathcal{F}_2(c) = \text{span}_{\mathbb{R}} \{c_n(\mu_0, \dots, \mu_{n-1}) : \mu_i \in \mathbb{R}^m, 0 \leq i \leq n-1, n \geq 0\}.$$

Let $\mathcal{G}_2(c)$ be the subspace of operators spanned by $F_{c_n(\mu_0, \mu_1, \dots, \mu_{n-1})}$ for all n and all choices of μ_0, \dots, μ_{n-1} . Then $\mathcal{F}_2(c)$ is isomorphic to $\mathcal{G}_2(c)$ (cf. [30]).

Clearly, for any power series c , $\mathcal{F}_2(c) \subseteq \mathcal{F}_1(c)$, since, for each integer n , $c_n(X_0, \dots, X_{n-1})$ is a polynomial on the X_i 's with coefficients belonging to $\mathcal{F}_1(c)$. A less trivial conclusion is that $\mathcal{F}_1(c) \subseteq \mathcal{F}_2(c)$. The following is an outline of the proof of this conclusion; for the detailed proof, refer to [30].

For any fixed positive integers k and i_1, i_2, \dots, i_q such that

$$1 \leq i_1 \leq i_2 \leq \dots \leq i_q \leq m,$$

let

$$S^k(i_1, i_2, \dots, i_q) = \left\{ \sigma(\underbrace{0, \dots, 0}_k, i_1, i_2, \dots, i_q) : \sigma \in S_n \right\},$$

where $n = k + q$ and S_n is the permutation group on a set of n elements. Let

$$T_k(i_1, i_2, \dots, i_q) = \left\{ w = \eta_{l_1} \eta_{l_2} \dots \eta_{l_n} : (l_1, \dots, l_n) \in S^k(i_1, i_2, \dots, i_q) \right\}$$

and order the elements of $T_k(i_1, i_2, \dots, i_q)$ as W_1, W_2, \dots, W_r . Then, for any j_1, \dots, j_q given,

$$\begin{aligned} \Upsilon_{i_1 \dots i_q}^{j_1 \dots j_q}(k) &:= \Gamma_{i_1 \dots i_q}^{j_1 \dots j_q}(j_1 + \dots + j_q + k + q) \\ &= \eta_0^{(k)} \sqcup \eta_{i_1} X^{(j_1)} \sqcup \eta_{i_2} X^{(j_2)} \sqcup \dots \sqcup \eta_{i_q} X^{(j_q)} \Big|_{X=1} \end{aligned}$$

is a linear combination of the elements in $T_k(i_1, i_2, \dots, i_q)$. We now define

$$\Delta_k(i_1, \dots, i_q) = \left\{ \Upsilon_{i_1 \dots i_q}^{j_1 \dots j_q}(k) : j_s \geq 0, 1 \leq s \leq q \right\}.$$

Our conclusion can be proved by showing that every element of $T_k(i_1, i_2, \dots, i_q)$ is a linear combination of elements in $\Delta_k(i_1, i_2, \dots, i_q)$ for any i_1, \dots, i_q and k given.

For each fixed k and q and fixed i_1, i_2, \dots, i_q , we order the elements of $\Delta_k(i_1, i_2, \dots, i_q)$ as Q_1, Q_2, \dots . Then, for each Q_i , there exist $a_{ij}, j = 1, \dots, r$ such that

$$Q_i = \sum_{j=1}^r a_{ij} W_j.$$

Let A be the matrix of r columns and infinitely many rows whose (i, j) th entry is a_{ij} ; i.e., $A = (a_{ij})$.

We claim that A is of full column rank in the sense that there is no nonzero vector $v \in \mathbb{R}^r$ such that $Av = 0$. Suppose that there is some $v \neq 0$ such that $Av = 0$. Let a be the polynomial defined by

$$a = v_1 W_1 + v_2 W_2 + \cdots + v_r W_r,$$

where v_i is the i th component of v . Then, for any $w \in P^*$,

$$\langle w^{-1}a, \phi \rangle \neq 0$$

if and only if $w = W_i$ for some i . Hence

$$(20) \quad \langle \psi_a \left(\Upsilon_{s_1 \dots s_p}^{j_1 \dots j_p}(l) \right), \phi \rangle = 0$$

if $l \neq k$, $p \neq q$, or $s_t \neq i_t$ for some t . In the other words, (20) holds if

$$\Upsilon_{s_1 \dots s_p} j_1 \dots j_p(k) \notin \Delta_k(i_1, i_2, \dots, i_q).$$

For $Q_i \in \Delta_k(i_1, i_2, \dots, i_q)$, we have that

$$\langle \psi_a(Q_i), \phi \rangle = \sum_{j=1}^r a_{ij} \langle W_j^{-1}a, \phi \rangle = \sum_{j=1}^r a_{ij} \langle a, W_j \rangle = \sum_{j=1}^r a_{ij} v_j.$$

By assumption, however, $\sum a_{ij} v_j = 0$ for any i . Therefore (20) holds for any choice of s_1, \dots, s_p , j_1, \dots, j_p , and any l . It then follows directly from the definition of $a_n(X_0, \dots, X_{n-1})$ that

$$(21) \quad \langle a_n(\mu_0, \mu_1, \dots, \mu_{n-1}), \phi \rangle = 0$$

for any n and any value of μ_0, \dots, μ_{n-1} , which, by (17), implies that

$$\frac{d^l}{dt^l} F_a[u](0) = F_{a_n(\mu_0, \dots, \mu_{n-1})}[u](0) = \langle a_n(\mu_0, \mu_1, \dots, \mu_{n-1}), \phi \rangle$$

for any analytic control u . Thus $F_a[u] \equiv 0$ for any analytic control. It then follows from the continuity of F_a and the density property of analytic controls in L_1 controls that $F_a \equiv 0$, which in turn implies that $a = 0$, a contradiction to the assumption that $v \neq 0$. Hence A is of full column rank.

It is easy to see that there exists some submatrix A_1 of A with finitely many rows such that A_1 is full column rank, which implies that each W_i is a linear combination of finitely many Q_j 's.

The above discussion shows the following conclusion.

THEOREM 3.1. *For any power series c , $\mathcal{F}_1(c) = \mathcal{F}_2(c)$.*

4. i/o equations. In this section, we study high-order differential equations satisfied by inputs and outputs arising from i/o operators. To perform this study, we find it useful to introduce the algebraic concepts of observation algebra and observation field corresponding to any given series c .

The *observation algebra* $\mathcal{A}_2(c)$ is defined as the \mathbb{R} -algebra generated by the elements of $\mathcal{F}_2(c)$. By Lemma 2.2, $\mathcal{A}_2(c)$ is an integral domain; so its quotient field is well defined; we define the *observation field* of c as this quotient field. We see later that elementary properties of these algebraic objects serve to characterize the existence of i/o equations.

4.1. Definitions. By an *algebraic i/o equation of order k* , we mean an equation of the type

$$(22) \quad P\left(u(t), \dots, u^{(k)}(t), y(t), \dots, y^{(k)}(t)\right) = 0,$$

where

$$P \in \mathbb{R}[S_0, \dots, S_k, L_0, \dots, L_k]$$

is a polynomial nontrivial in L_k , and S_i denotes the set of m variables (S_{1i}, \dots, S_{mi}) .

DEFINITION 4.1. We say that a polynomial P as above is

(a) *rational* when

$$(23) \quad \begin{aligned} &P(S_0, \dots, S_k, L_0, \dots, L_k) \\ &= P_0(S_0, \dots, S_{k-1}, L_0, \dots, L_{k-1}) L_k + P_1(S_0, \dots, S_k, L_0, \dots, L_{k-1}); \end{aligned}$$

(b) *recursive* when

$$(24) \quad \begin{aligned} &P(S_0, \dots, S_k, L_0, \dots, L_k) \\ &= P_0(S_0, \dots, S_{k-1}) L_k + P_1(S_0, \dots, S_k, L_0, \dots, L_{k-1}). \end{aligned}$$

DEFINITION 4.2. Assume that c is a convergent power series. We say that the i/o operator F_c satisfies an *algebraic i/o equation* (22) if (22) holds for every possible C^k i/o pair

$$(u(t), y(t)) := (u(t), F_c[u](t))$$

of F_c for all $t \in [0, T]$ and for any T admissible for c . In that case, (22) is called an *i/o equation of F_c* .

An i/o operator F_c satisfies a *rational i/o equation* if P can be chosen rational, so that $P_0 = 0$ is *not* an i/o equation of F_c ; in another words, there exists some i/o pair (u, y) of F_c such that

$$(25) \quad P_0(u(t), u'(t), \dots, u^{(k)}(t), y(t), y'(t), \dots, y^{(k-1)}(t)) \neq 0$$

for some t . An i/o operator F_c satisfies a *recursive equation* if there is some such equation for which P is recursive.

The following lemma was proved in [28]; a detailed proof in the more general analytic case is given in [29].

LEMMA 4.3. F_c satisfies the i/o equation (22) if and only if

$$(26) \quad P\left(\mu_0, \dots, \mu_k, F_c, F_{c_1(\mu_0)}, \dots, F_{c_k(\mu_0, \dots, \mu_{k-1})}\right) = 0$$

for any $\mu_0, \mu_1, \dots, \mu_k \in \mathbb{R}^m$.

4.2. Properties of i/o equations. We now introduce the field

$$K = \mathbb{R}(\{S_{ij}, i = 1, \dots, m, j \geq 1\})$$

obtained by adjoining the indeterminates S_{ij} to \mathbb{R} . Let $\mathcal{F}^K, \mathcal{A}^K$ be the K -space and K -algebra generated by $c_n(S_0, \dots, S_{n-1})$ for all n . Let \mathcal{Q}^K be the quotient field of \mathcal{A}^K . Note that the field \mathcal{Q}^K is defined, since \mathcal{A}^K is an integral domain. The reason for this is essentially because \mathcal{A}^K can be naturally identified to the tensor product $\mathcal{A}_2 \otimes K$.

LEMMA 4.4. Let F_c be the i/o operator corresponding to the series c . The following properties then hold:

(a) If F_c satisfies a recursive i/o equation, then \mathcal{A}^K is a finitely generated K -algebra.

(b) If F_c satisfies an algebraic i/o equation, then \mathcal{Q}^K is a finitely generated field extension of K .

Proof. Consider $\hat{\mathcal{A}}^K$, the K -algebra generated by $F_{c_n(s_0, \dots, s_{n-1})}$ for all n . The assignment $\psi : c_n(\mu_0, \dots, \mu_{n-1}) \mapsto F_{c_n(\mu_0, \dots, \mu_{n-1})}$ is an isomorphism from $\mathcal{A}_2(c)$ onto $\hat{\mathcal{A}}_2(c)$, the \mathbb{R} -algebra generated by $F_{c_n(\mu_0, \dots, \mu_{n-1})}$. Thus ψ induces an isomorphism from \mathcal{A}^K onto $\hat{\mathcal{A}}^K$. Consequently, $\hat{\mathcal{Q}}^K$, the quotient field of $\hat{\mathcal{A}}^K$, is isomorphic to \mathcal{Q}^K . We prove conclusion (b) by showing that $\hat{\mathcal{Q}}^K$ is a finitely generated field extension of K , when F_c satisfies some algebraic equation.

It is easy to show, by taking the derivative with respect to time t on both sides of an algebraic i/o equation, that existence of an algebraic i/o equation for F_c implies that F_c also satisfies a rational i/o equation. Thus

$$(27) \quad \begin{aligned} &P_0(u(t), \dots, u^{(k)}(t), y(t), \dots, y^{(k-1)}(t)) y^{(k)}(t) \\ &= -P_1(u(t), \dots, u^{(k)}(t), y(t), \dots, y^{(k-1)}(t)), \end{aligned}$$

for some polynomials P_0 and P_1 , where $P_0 = 0$ is not an i/o equation of F_c . (See [28] for details, as well as [29] for an analogous result for analytic i/o equations.) By Lemma 4.3, we know that

$$\begin{aligned} &P_0(S_0, \dots, S_{k-1}, F_c, \dots, F_{c_{k-1}(S_0, \dots, S_{k-2})}) F_{c_k(S_0, \dots, S_{k-1})} \\ &= -P_1(S_0, \dots, S_k, F_c, \dots, F_{c_{k-1}(S_0, \dots, S_{k-2})}). \end{aligned}$$

Note that, since $P_0 = 0$ is not an i/o equation of F_c , there must exist some vector $(\mu_0, \dots, \mu_{k-1})$ such that

$$P_0(\mu_0, \dots, \mu_{k-1}, F_c, \dots, F_{c_{k-1}(\mu_0, \dots, \mu_{k-2})}) \neq 0,$$

which, in turn, implies that

$$P_0(S_0, \dots, S_{k-1}, F_c, \dots, F_{c_{k-1}(S_0, \dots, S_{k-2})}) \neq 0$$

as a polynomial in S_0, \dots, S_{k-1} . It follows from this discussion that

$$F_{c_k(S_0, \dots, S_{k-1})} \in \hat{\mathcal{Q}}_{k-1}^K,$$

where $\hat{\mathcal{Q}}_r^K$ denotes the field obtained by adjoining $F_c, F_{c_1(S_0)}, \dots, F_{c_r(S_0, \dots, S_{r-1})}$ to K .

Taking the derivative with respect to t on both sides of (27), we get that

$$(28) \quad \begin{aligned} &P_0(u(t), \dots, u^{(k)}(t), y(t), \dots, y^{(k-1)}(t)) y^{(k+1)}(t) \\ &= P_2(u(t), \dots, u^{(k+r)}(t), y(t), \dots, y^{(k+r-1)}(t)), \end{aligned}$$

where P_2 is some polynomial. By using the same argument as before, we show that

$$F_{c_{k+1}(S_0, \dots, S_k)} \in \hat{\mathcal{Q}}_k^K \subset \hat{\mathcal{Q}}_{k-1}^K.$$

By induction, we show that $\hat{\mathcal{Q}}^K = \hat{\mathcal{Q}}_{k-1}^K$. Since $\hat{\mathcal{Q}}_{k-1}^K$ is a finitely generated field extension of K —the generators are the coefficients of S_{ij} $i = 1, \dots, m$; $j = 0, 1, \dots, k-2$, in $F_c, F_{c_1}, \dots, F_{c_{k-1}}$ —we get the conclusion that \mathcal{Q}^K is also a finitely generated field extension of K . This completes the proof of (b); property (a) is proved in a similar fashion. \square

LEMMA 4.5. Let F_c be the i/o operator corresponding to the series c . The following properties then hold:

(a) If \mathcal{A}^K is a finitely generated K -algebra, then $\mathcal{A}_2(c)$ is a finitely generated \mathbb{R} -algebra;

(b) If \mathcal{Q}^K is a finitely generated field extension of K , then $\mathcal{Q}_2(c)$ is a finitely generated field extension of \mathbb{R} .

Proof. Again, we only provide the proof for part (b). Part (a) can be proved similarly.

Assume that \mathcal{Q}^K is a finitely generated field extension of K . Then there exists some $n > 0$, so that, for any $r \geq 0$, there exist two polynomials Q_0, Q_1 over K with

$$Q_0(c_0, c_1(S_1), \dots, c_{n-1}(S_0, \dots, S_{n-2})) \neq 0$$

such that

$$\begin{aligned} & Q_0(c_0, c_1(S_0), \dots, c_{n-1}(S_0, \dots, S_{n-1})) c_{n+r}(S_0, \dots, S_{n+r-1}) \\ &= Q_1(c_0, c_1(S_0), \dots, c_{n-1}(S_0, S_1, \dots, S_{n-2})). \end{aligned}$$

After clearing denominators and eliminating extra μ_j 's, we have an equation

$$\begin{aligned} & P_0(S_0, \dots, S_{n+r-1}, c_0, c_1(S_0), \dots, c_{n-1}(S_0, \dots, S_{n-2})) c_{n+r}(S_0, \dots, S_{n+r-1}) \\ &= P_1(S_0, \dots, S_{n+r-1}, c_0, c_1(S_0), \dots, c_{n-1}(S_0, \dots, S_{n-2})) \end{aligned}$$

with

$$P_0(S_0, \dots, S_{n+r-1}, c_0, c_1(S_0), \dots, c_{n-1}(S_0, \dots, S_{n-2})) \neq 0,$$

which implies that there exists some $(\mu_0, \dots, \mu_{n+r-1})$ so that

$$P_0(\mu_0, \dots, \mu_{n+r-1}, c_0, c_1(\mu_0), \dots, c_{n-1}(\mu_0, \dots, \mu_{n-2})) \neq 0,$$

or, equivalently,

$$P_0(\mu_0, \dots, \mu_{n+r-1}, F_c, F_{c_1(\mu_0)}, \dots, F_{c_{n-1}(\mu_0, \dots, \mu_{n-2})}) \neq 0.$$

This is an equation involving operators. It means that there exists some $u \in \mathcal{V}_T$, where T is admissible to c , and t such that

$$P_0(\mu_0, \dots, \mu_{n+r-1}, F_c[u](t), \dots, F_{c_{n-1}(\mu_0, \dots, \mu_{n-2})}[u](t)) \neq 0.$$

It follows from the fact that

$$P_0(\mu_0, \dots, \mu_{n+r-1}, F_c[u](t), \dots, F_{c_{n-1}(\mu_0, \dots, \mu_{n-2})}[u](t))$$

is a polynomial in $\mu_0, \dots, \mu_{n+r-1}$; the set

$$\Omega_1 := \left\{ \mu^{n+r-1} : P_0(\mu^{(n+r-1)}, F_c[u](t), \dots, F_{c_{n-1}(\mu^{n-2})}[u](t)) \neq 0 \right\}$$

is dense in $\mathbb{R}^{m \times (n+r)}$, where $\mu^l = (\mu_0, \dots, \mu_l)$ for any l . Define

$$\Omega = \{ \mu^{n+r-1} : P_0(\mu^{n+r-1}, c_0, \dots, c_{n-1}(\mu^{n-2})) \neq 0 \}.$$

Then $\Omega_1 \subseteq \Omega$. Thus Ω is dense in \mathbb{R}^{n+r} .

Clearly, if $\mu^{n+r-1} \in \Omega$, then $F_{c_{n+r}(\mu^{n+r-1})} \in \mathcal{T}_{n-1}$, the field obtained by adjoining all the coefficients of X_{ij} in $c_p(X_1, \dots, X_{p-1})$ for $p \leq n-1$ to \mathbb{R} . Applying Lemma 12.11 in [21], we see that $F_{c_{n+r}(\mu^{n+r-1})} \in \mathcal{T}_{n-1}$ for any $\mu^{n+r-1} \in \mathbb{R}^{n+r}$. Since r can be chosen arbitrarily, it follows that $\mathcal{Q}_2(c) = \mathcal{T}_{n-1}$, from which it follows that $\mathcal{Q}_2(c)$ is a finitely generated field extension of \mathbb{R} . \square

Combining Lemmas 4.4 and 4.5, we get the main result of this section shown below.

THEOREM 4.6. *Let F_c be the i/o operator corresponding to the series c . The following properties then hold:*

(a) *If F_c satisfies a recursive i/o equation, then $\mathcal{A}_2(c)$ is a finitely generated \mathbb{R} -algebra;*

(b) *If F_c satisfies an algebraic i/o equation, then $\mathcal{Q}_2(c)$ is a finitely generated field extension of \mathbb{R} .*

Remark 4.7. Generally, a field extension over \mathbb{R} with finite transcendence degree is not necessarily a finitely generated field extension of \mathbb{R} . By using Theorem 4.6, however, we can show that if the transcendence degree of $\mathcal{Q}_2(c)$ is finite, then it follows that $\mathcal{Q}_2(c)$ is a finitely generated field extension of \mathbb{R} . The reasoning is as follows. Assume that $\text{trdeg}_{\mathbb{R}} \mathcal{Q}_2(c) < \infty$, where $\text{trdeg}_{\mathcal{K}} \mathcal{Q}$ denotes the transcendence degree of \mathcal{Q} over \mathcal{K} for any fields \mathcal{Q} and \mathcal{K} . Now let \mathcal{L}_n be the set of all the coefficients of $c_n(S_0, \dots, S_{n-1})$, seen as a polynomial in S_0, \dots, S_{n-1} over \mathcal{S} , the ring of all series. Let $\mathcal{L} = \bigcup_n \mathcal{L}_n$. Then $\mathcal{Q}_2(c) = \mathbb{R}(\mathcal{L})$. On the other hand, $\mathcal{Q}^K = K(\mathcal{L})$. Therefore $\text{trdeg}_{\mathbb{R}} \mathcal{Q}_2(c) < \infty$ implies that

$$(29) \quad \text{trdeg}_K \mathcal{Q}^K < \infty.$$

If (29) holds, then there exists some n such that

$$c, c_1(S_0), \dots, c_n(S_0, \dots, S_{n-1})$$

are algebraically dependent over K ; i.e., there exists some polynomial P over K such that

$$P(c, c_1(S_0), \dots, c_n(S_0, \dots, S_{n-1})) = 0.$$

After clearing denominators and eliminating the extra S_{ij} , we get the following equation:

$$(30) \quad Q(S_0, \dots, S_k, c, c_1(S_0), \dots, c_n(S_0, \dots, S_{n-1})) = 0.$$

Note that if a convergent series c satisfies (30), then (30) is an algebraic i/o equation of F_c , which, by Theorem 4.6, implies that $\mathcal{Q}_2(c)$ is a finitely generated field extension of \mathbb{R} .

5. Realizability. We wish to study realization by “rational” systems, such as those studied in Bartosiewicz [1]. However, the question of possible poles in the right-hand side of the equation is very delicate, and it seems better, instead, to study a “singular” polynomial model, as we do next.

Just as i/o equations prove to be related to the structure of $\mathcal{A}_2(c)$ and $\mathcal{Q}_2(c)$, realizability forces the study of the observation algebra and observation field corresponding to the other type of observation space $\mathcal{F}_1(c)$. For a given power series c , we associate with it an *observation algebra* $\mathcal{A}_1(c)$ defined as the \mathbb{R} -algebra generated by the elements of $\mathcal{F}_1(c)$, and associate with it an *observation field* $\mathcal{Q}_1(c)$ defined

as the quotient field of $\mathcal{A}_1(c)$. Again, we know that $\mathcal{Q}_1(c)$ is defined, since $\mathcal{A}_1(c)$ is an integral domain. The result is, because of previous results, that $\mathcal{A}_1 = \mathcal{A}_2$ and $\mathcal{Q}_1 = \mathcal{Q}_2$ for every c , but the facts in this section do not depend on the equality. They are more readily understood in terms of \mathcal{A}_1 and \mathcal{Q}_1 .

DEFINITION 5.1. Suppose that c is a convergent series and T is admissible for c . The i/o operator F_c is *realizable* by a *singular polynomial state-space system*

$$\Sigma = ((g_0, \dots, g_m), x_0, q, h)$$

if there exists an integer n , some $x_0 \in \mathbb{R}^n$, polynomial vector fields g_0, g_1, \dots, g_m on \mathbb{R}^n , and two polynomial functions $q, h : \mathbb{R}^n \rightarrow \mathbb{R}$ such that the following properties hold:

(a) For each $u \in \mathcal{V}_T$ and $y = F_c[u]$, there is some absolutely continuous function $x(\cdot)$ defined on $[0, T]$ and satisfying $x(0) = x_0$ such that

$$q(x(t))x'(t) = g_0(x(t)) + \sum_{i=1}^m u_i(t)g_i(x(t))$$

for almost all $t \in [0, T]$, and $y(t) = h(x(t))$ for all $t \in [0, T]$.

(b) The solution $x(\cdot)$ in part (a) is of class \mathcal{C}^ω if u is of class \mathcal{C}^ω , and $x(\cdot)$ is of class \mathcal{C}^{k+1} if u is of class \mathcal{C}^k .

(c) There holds the following *regularity* condition: There exists some set Ω of analytic inputs that is dense in $\mathcal{C}^\infty[0, T]$ (with respect to the Whitney topology) such that for any $u \in \mathcal{V}_T \cap \Omega^m$, there exists some \mathcal{C}^ω solution $x(\cdot)$ as in (a), so that $q(x(\cdot)) \neq 0$.

If F_c can be realized by a singular polynomial system with $q(x) \equiv 1$, we say that F_c is realizable by a *polynomial* state-space system.

It can be seen from Definition 5.1 that, if $q(x) \neq 0$ for any $x \in \mathbb{R}^n$, then F_c is realizable (globally) by an analytic system in the usual sense. If $q(x_0) \neq 0$, then F_c is realizable locally by an analytic system.

The nondegeneracy condition proves to be equivalent (as shown in the proof below) to the fact that, for “almost every” i/o pair, $q(x(t)) \neq 0$ for almost every t . It could happen, however, that q vanishes along some trajectories.

The following theorem is the main result of this section. It constitutes a converse to Theorem 4.6, but in terms of different algebraic objects.

THEOREM 5.2. *Let F_c be the i/o operator corresponding to the series c . The following properties then hold:*

(a) *If $\mathcal{A}_1(c)$ is a finitely generated \mathbb{R} -algebra, then F_c is realizable by a polynomial system;*

(b) *If $\mathcal{Q}_1(c)$ is a finitely generated field extension of \mathbb{R} , then F_c is realizable by a singular polynomial system.*

Proof. As in the proof of Theorem 4.6, we only provide proof of part (b). Part (a) can be proved by the same argument without involving the regularity property.

Suppose that $\mathcal{Q}_1(c)$ is a finitely generated field extension of \mathbb{R} ; i.e., there exist some c_1, c_2, \dots, c_n such that

$$\mathcal{Q}_1(c) = \mathbb{R}(c_1, c_2, \dots, c_n).$$

Without loss of generality, we may assume that $c_i \in \mathcal{A}_1(c)$ for $i = 1, 2, \dots, n$ and $c_1 = c$. For each c_i and η_j , there exist some $q_{ij}, g_{ij} \in \mathbb{R}[X_1, X_2, \dots, X_n]$ such that

$$q_{ij}(c_1, c_2, \dots, c_n)(\eta_j^{-1}c) = g_{ij}(c_1, c_2, \dots, c_n),$$

for $i = 1, 2, \dots, n$, $j = 0, 1, \dots, m$, and $q_{ij}(c_1, c_2, \dots, c_n) \neq 0$. Without loss of generality, we may assume that $q_{ij} = q$ for all i, j . Otherwise, we may let

$$q(c_1, c_2, \dots, c_n) = \prod_{i,j} q_{ij}(c_1, c_2, \dots, c_n)$$

and change the g_{ij} accordingly. It follows from the fact that \mathcal{S} is an integral domain that

$$(31) \quad q(c_1, c_2, \dots, c_n) \neq 0.$$

For $j = 0, 1, \dots, m$, let $g_j = (g_{1j}, g_{2j}, \dots, g_{nj})'$, where “ $'$ ” denotes the transpose. Let $x_0 = (\langle c_1, \phi \rangle, \langle c_2, \phi \rangle, \dots, \langle c_n, \phi \rangle)'$ and $h(x) = x_1$. For $u \in \mathcal{V}_T$, let

$$(32) \quad x(t) = (F_{c_1}[u](t), F_{c_2}[u](t), \dots, F_{c_n}[u](t))'.$$

Then $x(0) = x_0$,

$$q(x(t))x'(t) = g_0(x(t)) + \sum_{j=1}^m u_j(t)g_j(x(t))$$

for almost all $t \in [0, T]$, and $y(t) = h(x(t))$. Thus the system

$$\begin{aligned} q(x)x' &= g_0(x) + \sum g_j(x)u_j, \\ x(0) &= x_0, \\ y &= h(x) \end{aligned}$$

realizes F_c if the regularity property of the system holds. To verify the regularity condition for this realization, let $d = q(c_1, c_2, \dots, c_n)$. Then $F_d \neq 0$. Note that polynomial controls are dense in \mathcal{V}_T with respect to the L_1 norm, and F_d is a continuous operator. Hence there is at least one polynomial control $p \in \mathbb{R}[t]$ such that $F_d[p] \neq 0$. It follows from the fact that, for any t , $F_d[p](t)$ depends analytically on the coefficients of t in $p(t)$ (cf. [28]) that $F_d[u] \neq 0$ for all polynomial controls u in a dense set of \mathcal{V}_T , which is the desired regularity property. \square

6. Main results. In this section we establish the equivalence between realizability and the existence of i/o equations. Recall that any convergent series c induces an i/o operator F_c on \mathcal{V}_T for which T is admissible for c . The following is our main result in this work.

THEOREM 6.1. *Assume that c is a convergent power series, let $T > 0$ be admissible for c , and let F_c be the i/o operator induced by c on \mathcal{V}_T . Then*

- (a) *The following statements are equivalent:*
 - (i) F_c satisfies an algebraic i/o equation;
 - (ii) F_c satisfies a rational i/o equation;
 - (iii) F_c is realizable by a singular polynomial system; and
- (b) F_c is realizable by a polynomial system if F_c satisfies a recursive i/o equation.

The realizability implications follow from Theorems 3.1, 4.6, and 5.2. The converses, i.e., the existence of equations assuming realizability, are quite straightforward exercises in elimination theory, and the details are given next.

LEMMA 6.2. *Assume that c is a convergent power series. Then F_c satisfies an algebraic i/o equation if F_c is realizable by a singular polynomial system.*

Proof. Assume that c is a convergent power series. We must prove that F_c satisfies some i/o equation

$$(33) \quad P(u(t), \dots, u^{(k)}(t), y(t), \dots, y^{(k)}(t)) = 0$$

valid for all \mathcal{C}^k i/o pairs (u, y) with $u \in \mathcal{V}_T$, and any T admissible for c . We henceforth fix such a T , and we assume that F_c is realized by the singular polynomial system

$$(34) \quad q(x)x' = g_0(x) + \sum_{j=0}^m u_j g_j(x), \quad x \in \mathbb{R}^n,$$

$$(35) \quad x(0) = x_0, \quad x_0 \in \mathbb{R}^n,$$

$$(36) \quad y = h(x), \quad y \in \mathbb{R}.$$

Assume for now that $q(x_0) \neq 0$. Then there exists some neighborhood \mathcal{N} of x_0 in \mathbb{R}^n such that $q(x) \neq 0$ for all $x \in \mathcal{N}$. Note that, on \mathcal{N} , (34) can be written as

$$(37) \quad x' = p_0(x) + \sum_{j=0}^m u_j p_j(x),$$

where $p_j = g_j/q$, for $j = 0, 1, \dots, m$.

Let $\varphi(t, x, u)$ denote the solution of (37) corresponding to the control u with the initial condition $x(0) = x$. Let $y_x(t) = h(\varphi(t, x, u))$. Then

$$y_x(0), y'_x(0), \dots, y_x^{(n)}(0)$$

are rational functions of x over the field of K , the field obtained by adjoining μ_{ij} ($i = 0, \dots, n-1$, $j = 0, \dots, m$) to \mathbb{R} . Since the transcendence degree of $K(x)$ over K is n , the $n+1$ rational functions $y_x(0), y'_x(0), \dots, y_x^{(n)}(0)$ are algebraically dependent over K ; i.e., there exists some nontrivial polynomial Q over K such that

$$Q(y_x(0), y'_x(0), \dots, y_x^{(n)}(0)) = 0.$$

Clearing the denominators in the coefficients (rational functions in the variables μ_0, \dots, μ_{n-1}), we obtain that

$$P(\mu_0, \dots, \mu_{n-1}, y_x(0), \dots, y_x^{(n)}(0)) = 0,$$

where $P \in \mathbb{R}[Y, \mu_0, \dots, \mu_{n-1}]$ is some polynomial over \mathbb{R} . Note here that P is nontrivial in Y , since Q is nontrivial.

Since P was chosen independent of the initial state x , it follows that, for any $u \in \mathcal{V}_T$, there exists some $\delta > 0$ such that

$$(38) \quad P(u(t), \dots, u^{(n-1)}(t), y(t), \dots, y^{(n)}(t)) = 0$$

for $t < \delta$. By principle of analytic continuation, (38) holds for all $t \in [0, T]$ and analytic controls in \mathcal{V}_T . Since analytic controls are dense in \mathcal{V}_T and F_c is continuous, (38) holds for all controls in \mathcal{V}_T .

Finally, we show how to overcome the restriction $q(x_0) \neq 0$. Assume now that $q(x_0) = 0$. Then, by definition, there exists a set Ω of analytic inputs in \mathcal{C}^∞ , open

dense with respect to the Whitney topology, so that, for each $u \in \Omega \cap \mathcal{V}_T$, there exists some analytic function $\varphi(t)$ satisfying (34) and (35) such that $q(\varphi(\cdot)) \neq 0$ and $F_c[u](t) = h(\varphi(t))$. It follows from analyticity that there exists some $\delta > 0$ such that $q(\varphi(t)) \neq 0$ for $t \in (0, \delta)$. From the previous argument, we see that $(u(t), y(t))$ satisfies (38) for any $t \in (0, \delta)$. Using analyticity again, we know that $(u(t), F_c[u](t))$ satisfies (38) for all $t \in [0, T]$.

Since Ω is dense in \mathcal{C}^∞ controls and \mathcal{C}^∞ controls are dense in \mathcal{C}^n controls with respect to the Whitney topology, it follows that (38) holds for all \mathcal{C}^n controls in \mathcal{V}_T . \square

Note that, in contrast to the cases of the rational i/o equation, the converse of part (b) does not hold in general; i.e, realizability by polynomials system does not necessarily imply the existence of a recursive i/o equation. This can be illustrated by the following example.

Example 6.3. Consider the following system:

$$(39) \quad \begin{aligned} x'_1 &= x_1 x_2, & x_1(0) &= x_{10} = 1; \\ x'_2 &= u, & x_2(0) &= x_{20} = 0; \\ y &= x_1. \end{aligned}$$

Then there exists some $T > 0$ such that, for all $u \in \mathcal{V}_T$, $y(t) = F_c[u](t)$, where c is given by

$$\langle c, \eta_{i_1} \eta_{i_2} \cdots \eta_{i_l} \rangle = L_{g_{i_1}} \cdots L_{g_{i_l}} L_{g_{i_1}} h(x_0),$$

where $g_0 = x_1 x_2 (\partial/\partial x_1)$, $g_1 = \partial/\partial x_2$, and $h(x) = x_1$ (cf. [13]). In the other words, F_c is realizable by the polynomial system (39).

To show that the operator F_c does not satisfy any recursive i/o equation, we must first establish the following fact. To a general analytic state space system

$$(40) \quad x' = g_0(x) + \sum_{i=1}^m g_i(x), \quad x \in \mathcal{M}, \quad y = h(x),$$

we associate an observation space F_1 defined as \mathbb{R} -space spanned by all the functions

$$L_{g_{i_1}} L_{g_{i_2}} \cdots L_{g_{i_k}} h(x), \quad k \geq 0, \quad 0 \leq i_1, i_2, \dots, i_k \leq m.$$

We define the *observation algebra* \mathcal{A} of (40) as the \mathbb{R} -algebra generated by the elements of F_1 .

For each $x_0 \in \mathcal{M}$, let c_h be the generating series defined by

$$(41) \quad \langle c_h, \eta_{i_1} \eta_{i_2} \cdots \eta_{i_l} \rangle = L_{g_{i_1}} \cdots L_{g_{i_l}} L_{g_{i_1}} h(x_0).$$

We say that system (40) is *accessible* at x_0 if, for any neighborhood \mathcal{B} of x_0 , there exists an open subset of \mathcal{U} of \mathcal{B} such that, for any $p \in \mathcal{U}$, there exist some $\tau \geq 0$ and some $u \in L_\infty^m[0, \tau]$ such that $x(\tau, x_0, u) = p$. The following lemma is provided in [28].

LEMMA 6.4. *Assume that the analytic system (40) is accessible at x_0 and that \mathcal{M} is connected. Let c_h be the series defined by (41). Then the observation algebra $\mathcal{A}_1(c_h)$ associated with c_h is isomorphic to the observation algebra \mathcal{A} associated with (40).*

System (34) is accessible at $x_0 = (1, 0)$ since the accessibility rank condition (see, for instance, [24]) holds, as follows:

$$\text{rank} \begin{pmatrix} g_0(x_0) & [g_0, g_1](x_0) \end{pmatrix} = 2.$$

If F_c would satisfy some recursive i/o equation, then the observation algebra $\mathcal{A}_2(c)$ would be finitely generated, which, by Lemma 6.4, would imply that \mathcal{A} is also finitely generated as an \mathbb{R} -algebra. This is false, however, as \mathcal{A} is the algebra generated by

$$x_1, x_1x_2, x_1x_2^2, \dots, x_1x_2^k, \dots \quad k \geq 0.$$

Thus F_c cannot satisfy any recursive i/o equation, even though it is realized by the polynomial system (39).

7. Families of i/o operators. In this section we study families of power series and i/o operators. Let Λ be an index set. We say that \underline{c} is a *family of power series* (parameterized by $\lambda \in \Lambda$) if $\underline{c} = \{c^\lambda : \lambda \in \Lambda\}$, where c^λ is a power series for each fixed λ . A family \underline{c} can also be viewed as a power series with coefficients belonging to a ring of functions from Λ to \mathbb{R} ; i.e. $\underline{c} = \sum \langle \underline{c}, \eta_\iota \rangle \eta_\iota$, where $\langle \underline{c}, \eta_\iota \rangle : \Lambda \rightarrow \mathbb{R}, \lambda \mapsto \langle c^\lambda, \eta_\iota \rangle$ is a function defined on Λ .

Thus we may treat families of power series as power series over some ring R . We use \mathcal{S}_R to denote the set of all power series over R . Then \mathcal{S}_R is a ring with “+” and “ \sqcup ” defined as the following:

$$\gamma \underline{c} + \underline{d} = \{\gamma c^\lambda + d^\lambda : \lambda \in \Lambda\},$$

$$\underline{c} \sqcup \lambda = \{\underline{c}^\lambda \sqcup \underline{d}^\lambda : \lambda \in \Lambda\},$$

for all $\underline{c}, \underline{d} \in \mathcal{S}_R, \gamma \in \mathbb{R}$.

Unlike the set \mathcal{S} of power series over \mathbb{R} , \mathcal{S}_R may not be an integral domain. This is because ring R may not be an integral domain. However, by following the same steps in the proof of Lemma 2.2, we can get the following conclusion.

LEMMA 7.1. *The ring \mathcal{S}_R is an integral domain if R is an integral domain.*

It follows from the principle of analytic continuation that any ring of analytic functions from a connected analytic manifold to \mathbb{R} is an integral domain. So we have the following fact.

COROLLARY 7.2. *If Λ is a connected analytic manifold and R is a ring of analytic functions from Λ to \mathbb{R} , then \mathcal{S}_R is an integral domain.*

DEFINITION 7.3. We say a family \underline{c} is a *convergent family* if

- (a) Each member of the family is convergent;
- (b) Λ is a topological space, $\langle c^\lambda, \eta_\iota \rangle$ depends on λ continuously, for each $\eta_\iota \in P^*$, and the constants K_λ, M_λ as in (8) can be chosen continuously depending on λ .

Since each convergent series induces an i/o operator, each convergent family c of power series induces a family of i/o operators $\{F_{c^\lambda} : \lambda \in \Lambda\}$, which we denote by $\mathbf{F}_{\underline{c}}$. The following result is provided in [28].

LEMMA 7.4. *Assume that \underline{c} is a convergent family. If T is admissible for c^{λ_0} , then T is admissible for c^λ for all λ in a small neighborhood of λ_0 , and $F_c^\lambda[u](t)$ depends (jointly) continuously on t and λ .*

7.1. Observation spaces for families of i/o operators. For a family \underline{c} of power series, we define $z^{-1}\underline{c}$ to be the family $\{z^{-1}c : \lambda \in \Lambda\}$, for any $z \in P^*$. For any $n \geq 0$, $\underline{c}_n(X_0, \dots, X_{n-1})$ is defined to be the family

$$\{c_n^\lambda(X_0, \dots, X_{n-1}) : \lambda \in \Lambda\},$$

where $X_i = (X_{i1}, \dots, X_{im})$ are m indeterminates over \mathbb{R} , $i \geq 0$.

As in the case of single power series, we associate to \underline{c} two types of observation spaces in the following way:

$$\tilde{\mathcal{F}}_1(\underline{c}) := \text{span}_{\mathbb{R}} \{ \alpha^{-1} \underline{c} : \alpha \in P^* \},$$

$$\tilde{\mathcal{F}}_2(\underline{c}) := \text{span}_{\mathbb{R}} \{ \underline{c}_n(\mu_0, \dots, \mu_{n-1}) : \mu_i \in \mathbb{R}^m, 0 \leq i \leq n-1, n \geq 0 \}.$$

Note here that the elements of $\tilde{\mathcal{F}}_1(\underline{c})$ and $\tilde{\mathcal{F}}_2(\underline{c})$ are families of series. For instance, if \underline{c} is given by

$$c^\lambda = \lambda^2 + 2\lambda\eta_0 - \lambda^3\eta_1, \quad \lambda \in \mathbb{R},$$

then $\tilde{\mathcal{F}}_1(\underline{c})$ is spanned by three elements: \underline{c} , 2λ , and λ^3 ; thus $\tilde{\mathcal{F}}_1(\underline{c})$ is a three-dimensional \mathbb{R} -space.

Treating families of series as single series over a ring and following the same steps in the proof of Theorem 3.1, we can obtain an analogue of Theorem 3.1 for families, shown in the following theorem.

THEOREM 7.5. *For any family \underline{c} of power series, $\tilde{\mathcal{F}}_1(\underline{c}) = \tilde{\mathcal{F}}_2(\underline{c})$.*

7.2. i/o equations for families of i/o operators. We say that a family $\mathbf{F}_{\underline{c}}$ satisfies an algebraic i/o equation of order k if there exists some polynomial $P \in \mathbb{R}[S_0, \dots, S_k, L_0, \dots, L_k]$, nontrivial in L_k such that

$$(42) \quad P(u(t), \dots, u^{(k)}(t), y(t), \dots, y^{(k)}(t)) = 0$$

is an i/o equation for F_{c^λ} for each $\lambda \in \Lambda$.

If (42) is recursive, then we say that $\mathbf{F}_{\underline{c}}$ satisfies a recursive equation. We say that (42) is a rational i/o equation for $\mathbf{F}_{\underline{c}}$ if

$$\begin{aligned} &P(S_0, \dots, S_k, L_0, \dots, L_k) \\ &= P_0(S_0, \dots, S_k, L_0, \dots, L_{k-1})L_k + P_1(S_0, \dots, S_k, L_0, \dots, L_{k-1}) \end{aligned}$$

for some polynomials P_0 and P_1 , and P_0 is not an i/o equation for $\mathbf{F}_{\underline{c}}$; i.e., there exists some $\lambda \in \Lambda$ and some i/o pair (u, y) of F_{c^λ} that does not satisfy (42).

For a family of generating series \underline{c} , we associate with it an observation algebra $\tilde{\mathcal{A}}_2(\underline{c})$ defined as the \mathbb{R} -algebra generated by the elements of $\tilde{\mathcal{F}}_2(\underline{c})$. Recall that $\tilde{\mathcal{F}}_2(\underline{c})$ is the \mathbb{R} -space generated by $\underline{c}_n(\mu_0, \dots, \mu_{n-1})$ for all n and all μ .

To define the observation field, we need the assumption that $\tilde{\mathcal{A}}_2(\underline{c})$ is an integral domain.

DEFINITION 7.6. We say that a convergent family $\underline{c} = \{c^\lambda : \lambda \in \Lambda\}$ is an *analytic* family if Λ is a connected analytic manifold and $\langle c^\lambda, \eta_\iota \rangle$ is an analytic function defined on Λ for all $\iota \in P^*$.

By Corollary 7.3, $\tilde{\mathcal{A}}_2(\underline{c})$ is an integral domain; therefore, its quotient field is well defined. For an analytic family \underline{c} , we define the observation field $\tilde{\mathcal{Q}}_2(\underline{c})$ of \underline{c} as the quotient field of $\tilde{\mathcal{A}}_2(\underline{c})$.

By using the same ideas used in §4, we get the following conclusion.

THEOREM 7.7. *Assume that \underline{c} is an analytic family of power series. Then*

- (a) $\tilde{\mathcal{A}}_2(\underline{c})$ is a finitely generated \mathbb{R} -algebra if $\tilde{\mathcal{F}}_{\underline{c}}$ satisfies a recursive i/o equation;
- (b) $\tilde{\mathcal{Q}}_2(\underline{c})$ is finitely generated field extension of \mathbb{R} if $\tilde{\mathcal{F}}_{\underline{c}}$ satisfies an algebraic i/o equation.

7.3. Realizability for families of i/o operators. DEFINITION 7.8. We say that a family $\mathbf{F}_{\underline{c}}$ of i/o operators is *realizable by a singular polynomial state space system*

$$\Sigma = ((g_0, g_1, \dots, g_m), X, q, h),$$

where g_0, g_1, \dots, g_m are polynomial vector fields of \mathbb{R}^n , X is a subset of \mathbb{R}^n , q and h are polynomial functions defined on \mathbb{R}^n , if the following properties hold:

(a) For each $\lambda \in \Lambda$ and each $u \in \mathcal{V}_{T_\lambda}$, where T_λ is admissible for c^λ , there exists some absolutely continuous function $x^\lambda(\cdot)$ defined on $[0, T]$ satisfying $x^\lambda(0) = x_0^\lambda$ for some $x_0^\lambda \in X$ such that

$$q(x^\lambda(t))(x^\lambda(t))' = g_0(x^\lambda(t)) + \sum_{j=1}^m g_j(x^\lambda(t))u_j(t)$$

for almost all $t \in [0, T]$, and

$$F_{c^\lambda}[u](t) = h(x^\lambda(t))$$

for all $t \in [0, T]$ and all $\lambda \in \Lambda$.

(b) The solution $x^\lambda(\cdot)$ in part (a) is of class \mathcal{C}^ω if u is of class \mathcal{C}^ω , and $x^\lambda(\cdot)$ is of class \mathcal{C}^{k+1} if u is of class \mathcal{C}^k .

(c) There holds the following *regularity* condition: There exists some open dense set Λ_1 of Λ such that, for $\lambda \in \Lambda_1$, there exists some set Ω_λ of analytic functions that is dense in $\mathcal{C}^\infty[0, T_\lambda]$ (with respect to Whitney topology) such that, for any $u \in \mathcal{V}_{T_\lambda} \cap \Omega_\lambda^m$, there exists some \mathcal{C}^ω solution $x^\lambda(\cdot)$ as in (a), so that $q(x^\lambda(\cdot)) \neq 0$. If $\mathbf{F}_{\underline{c}}$ can be realized by a singular polynomial system with

$$q(x) = 1 \quad \text{for all } x \in \mathbb{R}^n,$$

we say that $\mathbf{F}_{\underline{c}}$ is realizable by a polynomial system, and, if, in addition, the vector fields g_0, \dots, g_m are linear in x , then we say that $\mathbf{F}_{\underline{c}}$ is realizable by a bilinear system.

For an analytic family of power series \underline{c} , we associate with it an observation algebra $\tilde{\mathcal{A}}_1(\underline{c})$ defined as the \mathbb{R} -algebra generated by the elements of $\tilde{\mathcal{F}}_1(\underline{c})$ and an observation field $\tilde{\mathcal{Q}}_1(\underline{c})$ defined as the quotient field of $\tilde{\mathcal{A}}_1(\underline{c})$. Note here that the analyticity of the family implies that the quotient field of $\tilde{\mathcal{A}}_1(\underline{c})$ is well defined.

By using the same techniques used in §5, we get the following conclusion.

THEOREM 7.9. *Let \underline{c} be an analytic family of power series. Then*

(a) *The family of i/o operators $\mathbf{F}_{\underline{c}}$ is realizable by a polynomial system if $\tilde{\mathcal{A}}_1(\underline{c})$ is a finitely generated \mathbb{R} -algebra;*

(b) *The family of i/o operators $\mathbf{F}_{\underline{c}}$ is realizable by a singular polynomial system if $\tilde{\mathcal{Q}}_1(\underline{c})$ is a finitely generated field extension of \mathbb{R} .*

Combining all the results in this section, we see that the existence of i/o equations implies realizability. On the other hand, if $\mathbf{F}_{\underline{c}}$ is realizable by some singular polynomial system, then, by using approximation arguments, we can show that $\mathbf{F}_{\underline{c}}$ must satisfy some algebraic i/o equation. Hence we have the following theorem.

THEOREM 7.10. *Assume that \underline{c} is an analytic families of series. Then*

(a) *The following statements are equivalent:*

- (i) $\mathbf{F}_{\underline{c}}$ satisfies an algebraic i/o equation;
- (ii) $\mathbf{F}_{\underline{c}}$ satisfies a rational i/o equation;
- (iii) $\mathbf{F}_{\underline{c}}$ is realizable by a singular polynomial system; and

(b) \mathbf{F}_c is realizable by a polynomial system if \mathbf{F}_c satisfies a recursive i/o equation.

Remark 7.11. In the proofs of parts (a) of Theorems 7.9 and 7.10, we need not assume that $\tilde{\mathcal{A}}_1(c)$ and $\tilde{\mathcal{A}}_2(c)$ are integral domains. Hence part (b) of Theorem 7.10 also holds for continuous families; that is, for continuous families of operators, existence of recursive i/o equation implies realizability by polynomial systems.

8. Closing remarks. We envision our results being used as follows (the idea is very similar to that employed in the discrete case, and explored in some detail in [5]). If there are reasons to believe that the system producing the observed data is well posed, then an equation E may be fit to the data. We are *assured* that there is then a realization of the type to be considered, and we then try to find this realization. We are still very far from having constructive techniques for obtaining realizations; this is a major topic for further research involving symbolic computation. The following example illustrates the type of construction suggested by the proofs.

Consider the i/o equation

$$(43) \quad uy'' = y^2u^2 + y'u'$$

and assume that it is “well posed” in the sense mentioned above; that is, there is a Fliess operator $y = F_c[u]$ for which every pair $(u, F_c[u])$ satisfies the equation. Then we know that F_c can be realized by some polynomial state space system

$$(44) \quad x' = f(x) + g(x)u,$$

$$(45) \quad y = h(x)$$

with some fixed initial state. We now try to deduce what f , g , and h should be. We have that

$$\begin{aligned} y' &= L_f h(x) + L_g h(x)u, \\ y'' &= L_f^2 h(x) + (L_f L_g h(x) + L_g L_f h(x))u + L_g^2 h(x)u^2 + L_g h(x)u'. \end{aligned}$$

Substituting y, y', y'' into (43), we get the following formulas:

$$(46) \quad L_f h = 0,$$

$$(47) \quad L_f L_g h + L_g L_f h = h^2,$$

$$(48) \quad L_g^2 h = 0.$$

Formulas (46) and (47) suggest that $L_f^2 h = 0$ and $L_f L_g h = h^2$. Now let

$$z_1 = h(x), \quad z_2 = L_g h(x).$$

Then, along any trajectory $x(t)$ of (44),

$$\begin{aligned} z_1'(t) &= L_f h(x(t)) + L_g h(x(t))u(t) = z_2(t)u(t), \\ z_2'(t) &= L_f L_g h(x(t)) + L_g^2 h(x(t))u(t) = z_1(t)^2. \end{aligned}$$

Hence F_c can be realized by the following polynomial system:

$$z_1' = z_2 u, \quad z_2' = z_1^2, \quad y = z_1,$$

where the choice of initial state depends on additional data (such as the knowledge of $y(0)$ and $y'(0)$ for some nonzero control).

Of course, for practical applications, it is not clear when we would be justified in assuming wellposedness. We take the position, however, that postulating wellposedness is a far weaker assumption than assuming that the data was produced by a linear system, an assumption that itself underlies most applications of control theory.

Sometimes, we impose a “causality” constraint on i/o equations, requiring that the highest derivative of u be of lower order than derivatives of y . However, it is easy to see (cf. [28]) that, for i/o behaviors described by generating series, an equation of the type (1) always leads to an equation in which the highest order of derivative of inputs is lower than the highest order of derivative of outputs, i.e., an equation of the type

$$E\left(u(t), u'(t), u''(t), \dots, u^{(r-1)}(t), y(t), y'(t), y''(t), \dots, y^{(r)}(t)\right) = 0.$$

This is essentially a consequence of the fact that an i/o operator induced by a generating series must be causal in the sense that the k th-order derivatives of outputs do not depend on the k th-order derivatives of inputs.

Though nonsingular systems are preferred, we do not yet know if there is always a realization of that type (for nonrecursive equations). However, the analytic results in [29] can be applied to prove that about every singular point of the realization obtained here is another system, locally defined in terms of analytic functions, that realizes (locally) the desired behavior. The picture that emerges then is that, at least, we can cover the possibly singular part with local analytic realizations. In a computer simulation, this would be achieved by passing to a subroutine to deal with trajectories near this set.

As a final remark, we explain how this work relates to alternative foundations for systems theory recently proposed by various authors. We may consider the *behavior* $w(\cdot) = (u(\cdot), y(\cdot))$ associated to an i/o description. It has been proposed by [31] that we should formulate systems modeling without a priori distinctions between input and output signals. In these terms, an i/o equation takes the form

$$(49) \quad E\left(w(t), w'(t), w''(t), \dots, w^{(r)}(t)\right) = 0.$$

One of the central questions in [31] and related works is that of, in some sense, partitioning an abstract behavior $w(\cdot)$ into “inputs” and “outputs.” Once this task is achieved, however, and, provided that we may assume a suitable structure—in our case, the existence of a Fliess-operator relationship between inputs and outputs—it is still important to be able to relate an abstract equation such as (49) to realizability, and this is precisely what our result does. Similarly, the work [9] *defined* realizability by the requirement that outputs be differentially dependent on inputs; in other words, an equation such as (1) hold. We showed that this is basically the same as realizability in the more classical sense.

REFERENCES

- [1] Z. BARTOSIEWICZ, *Rational systems and observation fields*, Systems Control Lett., 9 (1987), pp. 379–386.
- [2] ———, *Minimal polynomial realizations*, Math. Control Signals Systems, 1 (1988), pp. 227–231.
- [3] G. CONTE, G. H. MOOG, AND A. PERDON, *Un théorème sur la représentation entrée-sortie d'un système non linéaire*, C. R. Acad. Sci. Paris, t. 307, Série I, 1988, pp. 363–366.
- [4] P. CROUCH AND F. LAMNABHI-LAGARRIGUE, *State space realizations of nonlinear systems defined by input-output differential equations*, in Proc. 8th Internat. Conf. Analysis Optimiz.

- Systems, Antibes, 1988, A. Bensoussan and J. L. Lions, eds., Springer-Verlag, Berlin, 1988, pp. 138–149.
- [5] H. DIAZ AND A. DESROCHERS, *Modeling of nonlinear discrete time systems from input-output data*, Automatica, 24 (1988), pp. 629–641.
 - [6] S. DIOP, *Elimination in control theory*, Math. Control, Signals Systems, 4 (1991), pp. 17–32.
 - [7] M. FLIESS, *Fonctionnelles causales non linéaires et indéterminées non commutatives*, Bull. Soc. Math. France, 109 (1981), pp. 3–40.
 - [8] ———, *Réalisation locale des systèmes non linéaires, algèbres de lie filtrées transitives et séries génératrices non commutatives*, Invent. Math., 71 (1983), pp. 521–537.
 - [9] ———, *Automatique et corps différentiels*, Forum Math., 1 (1989), pp. 227–238.
 - [10] M. FLIESS AND C. REUTENAUER, *Une application de l'algèbre différentielle aux systèmes réguliers (ou bilinéaires)*, in Analysis and Optimization of Systems, Lectures Notes in Control and Inform. Sci., A. Bensoussan and J. L. Lions, eds., Springer-Verlag, Berlin, 1982, pp. 99–107.
 - [11] M. FLIESS AND I. KUPKA, *A finiteness criterion for nonlinear input-output differential systems*, SIAM J. Control Optim., 21 (1983), pp. 721–728.
 - [12] S. T. GLAD, *Nonlinear state space and input output descriptions using differential polynomials*, in New Trends in Nonlinear Control Theory, J. Descusse, M. Fliess, A. Isidori, and M. Leborgne, eds., Springer-Verlag, Heidelberg, 1989, pp. 182–189.
 - [13] A. ISIDORI, *Nonlinear Control Systems: An Introduction*, Springer-Verlag, Berlin, 1985.
 - [14] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
 - [15] I. LEONTARITIS AND S.A. BILLINGS, *Input-output parametric models for nonlinear systems: Parts I and II*, Internat. J. Control, 41 (1985), pp. 303–344.
 - [16] J. T. LO, *Global bilinearization of systems with controls appearing linearly*, SIAM J. Control Optim., 13 (1975), pp. 879–885.
 - [17] N. H. MCCLAMROCH, *Singular systems of differential equations as dynamic models for constrained robot systems*, in Proc. IEEE Conf. Robotics and Automation, San Francisco, 1986, pp. 21–23.
 - [18] H. NIJMEIJER AND A. VAN DER SCHAFT, *Nonlinear Dynamical Control Systems*, Springer-Verlag, New York, 1990.
 - [19] R. REE, *Lie elements and an algebra associated with shuffles*, Ann. Math., 68 (1958), pp. 210–220.
 - [20] E. D. SONTAG, *On the internal realization of nonlinear behaviors*, in Dynamical Systems, A. Bednarek and L. Cesari, eds., Academic Press, New York, 1977, pp. 493–497.
 - [21] ———, *Polynomial Response Maps*, Springer-Verlag, Berlin, NY, 1979.
 - [22] ———, *Bilinear realizability is equivalent to existence of a singular affine differential i/o equation*, Systems Control Lett., 11 (1988), pp. 181–187.
 - [23] ———, *Mathematical Control Theory, Deterministic Finite Dimensional Systems*, Springer-Verlag, New York, 1990.
 - [24] H. J. SUSSMANN, *Lie brackets and real analyticity in control theory*, Math. Control Theory, Banach Center Publications, Warsaw, 14 (1985), pp. 515–542.
 - [25] M. E. SWEEDLER, *Hopf Algebras*, W. A. Benjamin, New York, 1969.
 - [26] A. V. VAN DER SCHAFT, *On realizations of nonlinear systems described by higher-order differential equations*, Math. Systems Theory, 19 (1987), pp. 239–275.
 - [27] ———, *Representing a nonlinear state space system as a set of higher-order differential equations in the inputs and outputs*, Systems Control Lett., 12 (1989), pp. 151–160.
 - [28] Y. WANG, *Algebraic Differential Equations and Nonlinear Control Systems*, Ph.D. thesis, Dept. of Mathematics, Rutgers University, New Brunswick, NJ, 1990.
 - [29] Y. WANG AND E. D. SONTAG, *Generating series and nonlinear systems: Analytic aspects, local realizability, and i/o representations*, Forum Math., 4 (1992), to appear.
 - [30] ———, *On two definitions of observation spaces*, Systems Control Lett., 13 (1989), pp. 279–289.
 - [31] J. C. WILLEMS, *Paradigms and puzzles in the theory of dynamical systems*, IEEE Trans. Automat. Control, TAC-36 (1991), pp. 259–294.
 - [32] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, Berlin, 2nd ed., 1979.

OPTIMAL SWITCHING AMONG SEVERAL BROWNIAN MOTIONS*

ROBERT J. VANDERBEI†

Abstract. For $i = 1, \dots, d$, let $B_{s_i}^i$ be a one-dimensional Brownian motion on the interval $[0, a_i]$ with absorption at the endpoints. At each instant in time, one must decide to run some subset of these d Brownian motions while holding the others fixed at their current state. The resulting process evolves in the rectangle $D = [0, a_1] \times \dots \times [0, a_d]$. If, at some instant, one decides to freeze all of the Brownian motions, then a reward is received in accordance with this final position. Two types of reward functions are considered.

First, it is assumed that the reward is zero everywhere in D , except along the d edges that correspond to the coordinate axes. Along these edges, it is given by C^3 strictly concave functions $\gamma_i(x_i)$, which are zero at the endpoints 0 and a_i of their domains. The optimal control for this problem has a simple description. Let

$$\Gamma_i(x_i) = - \int_0^{x_i} u \gamma_i''(u) du$$

and put

$$M_i = \{x \in D : \Gamma_i(x_i) = \max_j \Gamma_j(x_j)\}.$$

It is proved that the optimal control is: On M_i run any Brownian motion except the i th and stop the first time an edge is reached.

The second class of reward functions are assumed to be zero everywhere except on the facets of D that meet at the origin. On the i th such facet (i.e., where $x_i = 0$), the reward function is the product of $\gamma_j(x_j)$ for $j \neq i$. Put

$$N_i = \{x \in D : \Gamma_i(x_i) = \min_j \Gamma_j(x_j)\}.$$

The optimal control is: On N_i run the i th Brownian motion and stop when a facet of D is reached.

Key words. optimal control, Hamilton–Jacobi–Bellman equation, local time, Brownian motion, smooth fit

AMS (MOS) subject classifications. 60G40, 60J45, 31C10

1. Introduction. For $i = 1, \dots, d$, let $B^i = \{B_{s_i}^i, s_i \geq 0\}$ be a one-dimensional Brownian motion on the interval $[0, a_i]$ with absorption at the endpoints. We assume that B^i is adapted to a filtration $\mathcal{F}^i = \{\mathcal{F}_{s_i}^i, s_i \geq 0\}$ on the space Ω^i of continuous functions. Let $P_{x_i}^i$, $x_i \in [0, a_i]$ denote the probability measure associated with B^i starting at the point x_i , and let $\mathbf{E}_{x_i}^i$ denote the corresponding expectation operator. We assume that the filtration \mathcal{F}^i is complete with respect to every measure $P_{x_i}^i$, $x_i \in [0, a_i]$.

The problem that we study involves switching between these Brownian motions. We take as our sample space the product $\Omega = \Omega^1 \times \dots \times \Omega^d$ and let the Brownian motions be independent by putting $P_x = P_{x_1}^1 \times \dots \times P_{x_d}^d$ (\mathbf{E}_x denotes the corresponding expectation operator). A *switching strategy* T is a family of random d -tuples

$$(1) \quad T = \{T(t) = (T_1(t), \dots, T_d(t)), t \geq 0\},$$

* Received by the editors August 20, 1990; accepted for publication (in revised form) July 31, 1991.

† Program in Statistics and Operations Research, Princeton University, Princeton, New Jersey 08544 and AT&T Bell Laboratories, Murray Hill, New Jersey 07974.

satisfying

$$(2) \quad T(0) = (0, \dots, 0),$$

$$(3) \quad T_i(t) \text{ is increasing in } t \text{ for each } i,$$

$$(4) \quad T_1(t) + \dots + T_d(t) = t,$$

and

$$(5) \quad \{T_1(t) \leq s_1, \dots, T_d(t) \leq s_d\} \in \mathcal{F}_{s_1}^1 \times \dots \times \mathcal{F}_{s_d}^d.$$

The random variable $T_i(t)$ represents the amount of time the i th Brownian motion has been used up to time t . The interpretation of (4) is that, at time t , the total allocation of time between the d processes must equal t . Condition (5) says that the switching strategy must be nonanticipating. The *switched process* X^T is defined as

$$(6) \quad X^T(t) = B_{T(t)} = (B_{T_1(t)}^1, \dots, B_{T_d(t)}^d).$$

There are several possible criteria that may be optimized. Perhaps the most common is the accumulated discounted reward. In this case, we assume that each Brownian motion has a running reward function $r_i(x_i)$, and the problem then is to find the strategy T^* that attains the following supremum:

$$(7) \quad v(x) = \sup_T \mathbf{E}_x \int_0^\infty e^{-\lambda t} r(X^T(t)) \cdot dT(t),$$

where λ is a fixed positive constant, $r(x) = (r_1(x_1), \dots, r_d(x_d))$, and $r(X^T(t)) \cdot dT(t)$ represents the inner product between the vectors $r(X^T(t))$ and $dT(t)$. This problem was studied by Karatzas [2], Mandelbaum [3], and Dalang [1] as a continuous time generalization of Gittins' index theorem for Markov chains (see, e.g., Chap. 14 [6]). Assuming that each of the $r_i(x_i)$ are strictly increasing functions, they show that there exist functions $\Gamma_i(x_i)$ that determine the optimal strategy as follows:

$$(8) \quad T_i^*(t) \text{ increases only when } X^{T^*}(t) \in M_i,$$

where

$$(9) \quad M_i = \{x \in D : \Gamma_i(x_i) = \max_j \Gamma_j(x_j)\}.$$

This strategy is called a *follow-the-leader strategy*, since it runs process i when $\Gamma_i(x_i)$ is the largest of all the functions $\Gamma_j(x_j)$. The functions $\Gamma_j(x_j)$ are called *index functions*.

A different optimization criterion is considered in [4]. For $d = 2$, we study the problem of finding $T^*(t)$, which attains the following supremum:

$$(10) \quad v(x) = \sup_T \mathbf{E}_x f(X^T(\tau)),$$

where τ denotes the first time the switched process X^T exits a rectangle $D = (0, a_1) \times (0, a_2)$ and where f is a continuous pay-off function defined on the edges of D and *strongly concave* (i.e., twice continuously differentiable and strictly concave) or linear on each edge. In the case where f is zero except on the two edges that meet at the

origin, the optimal strategy has a simple description. Indeed, let $\gamma_i(x_i)$ denote the restriction of f to the x_i coordinate axis and put

$$(11) \quad \Gamma_i(x_i) = - \int_0^{x_i} u \gamma_i''(u) du = \gamma_i(x_i) - x_i \gamma_i'(x_i).$$

In terms of the sets M_i defined in (9), the optimal strategy satisfies the following condition:

$$T_i^*(t) \text{ increases only when } X^{T^*}(t) \notin M_i.$$

Hence the optimal strategy can be described as one that *follows the loser*.

The aim of this paper is to investigate how the above result generalizes to the case where $d > 2$. There are two possibilities. First, we could put concave data on the one-dimensional faces (i.e., edges) of D that meet at the origin. In this case, we let τ be the first hitting time of the set of edges of D .

THEOREM 1.1. *Let*

$$f(x) = \begin{cases} \gamma_1(x_1) & \text{if } x = (x_1, 0, \dots, 0) \\ \vdots & \\ \gamma_d(x_d) & \text{if } x = (0, \dots, 0, x_d) \\ 0 & \text{otherwise} \end{cases},$$

where each $\gamma_i(x_i)$ is C^3 , strictly concave, and vanishes at 0 and a_i . Optimal strategies exist. A strategy T^* is optimal if and only if

$$(12) \quad T_i^*(t) \text{ increases only when } X^{T^*}(t) \notin M_i$$

almost surely P_x , for all $x \in D$.

Hence optimal strategies are ones that follow anybody but the leader. For $d = 3$, the control regions are shown in Fig. 1 (in the case where the borders between the switching regions are planar).

Alternatively, we could put (certain types of) concave data on the codimension-one faces (i.e., facets) of D that meet at the origin. In this case, τ is assumed to be the first hitting time of a facet.

THEOREM 1.2. *Let*

$$f(x) = \begin{cases} \gamma_2(x_2) \cdots \gamma_d(x_d) & \text{if } x = (0, x_2, \dots, x_d), \\ \vdots & \\ \gamma_1(x_1) \cdots \gamma_{i-1}(x_{i-1}) \gamma_{i+1}(x_{i+1}) \cdots \gamma_d(x_d) & \text{if } x = (x_1, \dots, x_{i-1}, 0, \\ & \quad x_{i+1}, \dots, x_d), \\ \vdots & \\ \gamma_1(x_1) \cdots \gamma_{d-1}(x_{d-1}) & \text{if } x = (x_1, \dots, x_{d-1}, 0) \\ 0 & \text{otherwise} \end{cases}$$

where, for each i , $\gamma_i(x_i)$ is C^3 , strictly concave, and vanishes at 0 and a_i . There exists a unique (up to almost sure equivalence) optimal strategy. It satisfies the condition that

$$(13) \quad T_i^*(t) \text{ increases only when } X^{T^*}(t) \in N_i,$$

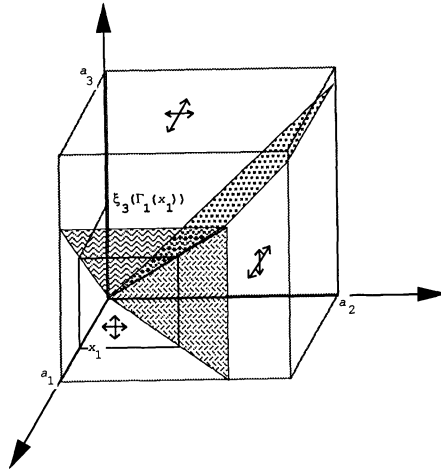


FIG. 1. Following anybody but the leader. Here we show the switching surfaces in the special case where each are planar. The arrows indicate which Brownian motions can be run in each region.

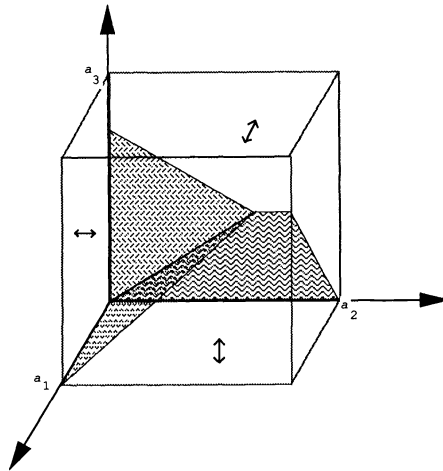


FIG. 2. Following the loser. Here we show the switching surfaces in the special case where each are planar. The arrows indicate which Brownian motions can be run in each region.

where

$$(14) \quad N_i = \{x \in D : \Gamma_i(x_i) = \min_j \Gamma_j(x_j)\}.$$

So, for facet data, the optimal strategy follows the loser. For $d = 3$, the control regions are shown in Fig. 2.

Remark. In Theorems 1 and 2, we assume that the boundary data is three times continuously differentiable. Two derivatives should suffice. However, we employ a change of variables in §§3 and 4, which necessitates our assumption of the existence of three derivatives. We believe that it should be possible to prove the results without using this change of variables, but the computations are more involved.

2. Probabilistic preliminaries. For the proofs of Theorems 1.1 and 1.2, we use the general theory of multiparameter processes. In this section, we review basic definitions and standard results. Our typical multiparameter process is a real-valued function of $(B_{s_1}^1, \dots, B_{s_d}^d)$, and so is always adapted to the multiparameter filtration $\mathcal{F} = \{\mathcal{F}_{s_1}^1 \times \dots \times \mathcal{F}_{s_d}^d : s_1 \geq 0, \dots, s_d \geq 0\}$.

A multiparameter process $\mathcal{M}_{s_1, \dots, s_d}$ is a *supermartingale* if it is adapted to \mathcal{F} , is integrable, and satisfies the supermartingale property: For every $x \in D$ and for all $s_1 \leq t_1, \dots, s_d \leq t_d$,

$$\mathbf{E}_x\{\mathcal{M}_{t_1, \dots, t_d} | \mathcal{F}_{s_1}^1 \times \dots \times \mathcal{F}_{s_d}^d\} \leq \mathcal{M}_{s_1, \dots, s_d}.$$

It is a *martingale* if the above inequality is replaced by equality.

Associated with any switching strategy $T(t)$, there is a one-parameter filtration $\mathcal{F}^T = \{\mathcal{F}_{T(t)} : t \geq 0\}$, where $\mathcal{F}_{T(t)}$ is defined as the σ -algebra containing all measurable sets C for which $C \cap \{T_1(t) \leq s_1, \dots, T_d(t) \leq s_d\} \in \mathcal{F}_{s_1}^1 \times \dots \times \mathcal{F}_{s_d}^d$ for all s_1, \dots, s_d . The switched process $X^T(t)$ is adapted to \mathcal{F}^T .

When we say that a multiparameter process is a martingale, we always mean that it is a martingale relative to \mathcal{F} . When we say that a one-parameter process, derived from a multiparameter process by following along a switching strategy $T(t)$, is a martingale, we mean that it is a martingale relative to \mathcal{F}^T .

A real-valued function defined on D is called *multiconcave* if it is concave in each component separately. It is *multilinear* if it is linear in each component separately.

PROPOSITION 2.1. *The following statements hold:*

1. If $\mathcal{M}_s = (\mathcal{M}_{s_1}^1, \dots, \mathcal{M}_{s_d}^d)$ is a multiparameter (super)martingale and $T(t)$ is a switching strategy, then $\mathcal{M}_{T(t)}$ is a (super)martingale;
2. If w is multilinear, then $w(X^T(t))$ is a martingale for any strategy $T(t)$;
3. If w is multiconcave, then $w(X^T(t))$ is a supermartingale for any strategy $T(t)$.

For $d = 2$, these results follow from Propositions 2.4 and 3.1 in [5]. The proofs given in [5] apply to $d > 2$, as well.

For the edge-data problem, let E denote the set of edges of D , and, for the facet-data problem, let E denote the set of facets of D . Then, in either case, τ is the first hitting time of E .

PROPOSITION 2.2. *Let w be a continuous, multiconcave function on D that agrees with f on E . If there exists a switching strategy $\tilde{T}(t)$ such that $w(B_{\tilde{T}(t \wedge \tau)})$ is a martingale, then w is the value function v defined in (10) and $\tilde{T}(t)$ is an optimal switching strategy.*

Proof. Appealing to (3) of Proposition 2.1 and the optional sampling theorem, we conclude that

$$(15) \quad w(x) \geq \mathbf{E}_x w(B_{T(\tau)}) = \mathbf{E}_x f(B_{T(\tau)})$$

for any switching strategy $T(t)$. Since $w(B_{\tilde{T}(t)})$ is a martingale, we see that

$$(16) \quad w(x) = \mathbf{E}_x w(B_{\tilde{T}(\tau)}) = \mathbf{E}_x f(B_{\tilde{T}(\tau)}).$$

From (15) and (16), we conclude that w is the value function and that $\tilde{T}(t)$ is an optimal switching strategy. \square

Now, to prove Theorems 1.1 and 1.2, two tasks remain. These are (i) to exhibit a function w that is continuous, multiconcave, and agrees with f on E , and (ii) to describe a switching strategy $\tilde{T}(t)$ for which $w(B_{\tilde{T}(t)})$ is a martingale.

For $w(B_{\tilde{T}(t)})$ to be a martingale, it seems to be necessary that $w(x)$ be linear in at least one component at every point $x \in D \setminus E$. Hence the function w should be a solution to the following nonlinear Dirichlet problem:

$$(17) \quad \max_{i: 0 < x_i < a_i} \frac{\partial^2 w}{\partial x_i^2}(x) = 0 \quad \text{for } x \in D \setminus E,$$

$$(18) \quad w(x) = f(x) \quad \text{for } x \in E.$$

In the next two sections, we construct twice continuously differentiable solutions to this differential equation.

Let w denote the solution to (17), (18), corresponding to either the “edge-data” problem or the “face-data” problem. We finish this section by constructing a switching strategy $\tilde{T}(t)$ for which $w(B_{\tilde{T}(t)})$ is a martingale. Consider any switching strategy $T(t)$. Since the functions x_i , and $x_i x_j$, for $j \neq i$, are multilinear, it follows from part 2 of Proposition 2.1 that $X_i^T(t)$ and $X_i^T(t)X_j^T(t)$ are martingales. Hence, for $t \geq 0$, the quadratic covariation between $X_i^T(t)$ and $X_j^T(t)$ vanishes:

$$(19) \quad \langle X_i^T, X_j^T \rangle_t = 0.$$

For each $i = 1, \dots, d$, the multiparameter process $(B_{s_i}^i)^2 - s_i$ is a multiparameter martingale, and so, by part 1 of Proposition 2.1, we see that $(X_i^T(t))^2 - T_i(t)$ is a martingale. Hence, for $t \geq 0$, the quadratic variation of $X_i^T(t)$ is given by

$$(20) \quad \langle X_i^T \rangle_t = T_i(t).$$

Since the function w constructed in either of the next two sections is C^2 , we can apply Itô’s formula, together with (19) and (20), to obtain that

$$(21) \quad \begin{aligned} w(X^T(t)) - w(X^T(0)) &= \sum_i \int_0^t \frac{\partial w}{\partial x_i}(X^T(s)) dX_i^T(s) \\ &+ \sum_i \int_0^t \frac{\partial^2 w}{\partial x_i^2}(X^T(s)) dT_i(s). \end{aligned}$$

Theorem 12 in [3] establishes the existence of a strategy $\tilde{T}(t)$ that “follows the smallest index function,”:

$$(22) \quad \tilde{T}_i(t) \text{ increases only when } X^{\tilde{T}}(t) \in N_i.$$

(In addition to existence, [3] also proves that the strategy is unique if the index processes $\Gamma_i(B_{s_i}^i)$ are simultaneously flat with probability zero. This condition is certainly met here for any pair of index processes and hence for any collection of them.) Whether considering the “edge-data” problem or the “facet-data” problem, in either case, the solution of (17), (18) constructed in the following sections has the property that

$$(23) \quad \frac{\partial^2 w}{\partial x_i^2}(x) = 0 \quad \text{for } x \in N_i$$

(this follows from the fact that $N_i \subset M_i^c$). Combining (22) and (23), we see that the second sum in (21) vanishes, and so $w(X^{\tilde{T}}(t))$ is a martingale.

3. Edge data. Let $w(x)$ denote a candidate for the value function $v(x)$ for the edge-data problem. Assuming that Theorem 1.1 correctly describes the optimal control regions, we see that $w(x)$ in the control region M_j should be linear in every component, except perhaps the j th. That is, the restriction of w to the intersection of the plane determined by a level set of x_j and the control set M_j should be multilinear. Hence we can “sweep out” $w(x)$ to the boundary of M_j . For $d = 2$ and for $x \in M_1$, this means that we can write

$$w(x_1, x_2) = \left(1 - \frac{x_2}{\xi_2(\Gamma_1(x_1))}\right) w(x_1, \xi_2(\Gamma_1(x_1))) + \frac{x_2}{\xi_2(\Gamma_1(x_1))} w(x_1, 0),$$

where

$$(24) \quad \xi_i(u) = \Gamma_i^{-1}(u \wedge \bar{u}_i)$$

and

$$(25) \quad \bar{u}_i = \Gamma_i(a_i).$$

Similarly, for $d = 3$ and $x \in M_1$, the formula becomes

$$\begin{aligned} w(x_1, x_2, x_3) = & \left(1 - \frac{x_2}{\xi_2(\Gamma_1(x_1))}\right) \left(1 - \frac{x_3}{\xi_3(\Gamma_1(x_1))}\right) w(x_1, 0, 0) \\ & + \frac{x_2}{\xi_2(\Gamma_1(x_1))} \left(1 - \frac{x_3}{\xi_3(\Gamma_1(x_1))}\right) w(x_1, \xi_2(\Gamma_1(x_1)), 0) \\ & + \left(1 - \frac{x_2}{\xi_2(\Gamma_1(x_1))}\right) \frac{x_3}{\xi_3(\Gamma_1(x_1))} w(x_1, 0, \xi_3(\Gamma_1(x_1))) \\ & + \frac{x_2}{\xi_2(\Gamma_1(x_1))} \frac{x_3}{\xi_3(\Gamma_1(x_1))} w(x_1, \xi_2(\Gamma_1(x_1)), \xi_3(\Gamma_1(x_1))). \end{aligned}$$

For the general formula, the notation can be streamlined by observing that the x_1 in the argument list for w can be written as $\xi_1(\Gamma_1(x_1))$. In general, for $x \in M_j$, $w(x_1, \dots, x_d)$ can be written as a sum over those subsets A of the set of indices $\{1, 2, \dots, d\}$ that contain j , as follows:

$$(26) \quad w(x_1, \dots, x_d) = \sum_{A: j \in A} \prod_{i \notin A} \left(1 - \frac{x_i}{\xi_i(\Gamma_j(x_j))}\right) \prod_{i \in A, i \neq j} \frac{x_i}{\xi_i(\Gamma_j(x_j))} w(T_A x_j),$$

where $T_A x_j$ is the d -dimensional point whose coordinates are given by

$$(27) \quad (T_A x_j)_i = \begin{cases} \xi_i(\Gamma_j(x_j)) & i \in A, \\ 0 & i \notin A. \end{cases}$$

Note that the product over $i \in A, i \neq j$ in (26) can actually be taken over all $i \in A$, since, for $i = j$, the factor $x_i/\xi_i(\Gamma_j(x_j))$ is just one.

Notations are greatly simplified if we change coordinates, so that $x_i = \xi_i(u_i)$ for each i . Then the function w becomes

$$(28) \quad \tilde{w}(u_1, \dots, u_d) = w(\xi_1(u_1), \dots, \xi_d(u_d)),$$

the domain D becomes $\tilde{D} = \{(u_1, \dots, u_d) : u_i \leq \bar{u}_i, \text{ for all } i\}$, and, for each i , the control region M_i becomes

$$(29) \quad \tilde{M}_i = \{(u_1, \dots, u_d) \in \tilde{D} : u_i = \max_j u_j\}.$$

Let \tilde{w}_j denote the restriction of \tilde{w} to \tilde{M}_j . Then

$$(30) \quad \tilde{w}_j(u_1, \dots, u_d) = \sum_{A: j \in A} \prod_{i \notin A} q_i(u_i, u_j) \prod_{i \in A} \xi_i(u_i) \theta_A(u_j),$$

where

$$(31) \quad q_i(u_i, u_j) = 1 - \frac{\xi_i(u_i)}{\xi_i(u_j)},$$

$$(32) \quad \theta_A(u) = \frac{\tilde{w}(\tilde{T}_A u)}{\prod_{i \in A} \xi_i(u)},$$

and

$$(33) \quad (\tilde{T}_A u)_i = \begin{cases} \bar{u}_i \wedge u & i \in A, \\ 0 & i \notin A. \end{cases}$$

As long as we ensure that \tilde{w}_j and \tilde{w}_k patch together smoothly along their border $\tilde{M}_j \cap \tilde{M}_k$, then it follows that \tilde{w} is smooth throughout \tilde{D} . Since the change of variables (28) involves twice continuously differentiable functions (which follows from our assumption that the functions γ_i are C^3), the smoothness of \tilde{w} in \tilde{D} translates back into the same smoothness of w in D (up to second order).

As we now show, stipulating first-order smoothness across $\tilde{M}_j \cap \tilde{M}_k$ forces θ_A to be a specific function for each A . Stipulating second-order smoothness forces the functions Γ_i to be as defined in (11).

Fix j, k with $j \neq k$. First, we note that the values of \tilde{w}_j and \tilde{w}_k agree along $\tilde{M}_j \cap \tilde{M}_k$:

$$\tilde{w}_j|_{u_j:=u_k:=u} = \sum_{A: j, k \in A} Q_A \Xi_A \theta_A(u) = \tilde{w}_k|_{u_j:=u_k:=u},$$

where Q_A and Ξ_A are abbreviations for the following expressions:

$$(34) \quad Q_A = \prod_{i \notin A} q_i(u_i, u),$$

$$(35) \quad \Xi_A = \prod_{i \in A} \xi_i(u_i),$$

and $\tilde{w}_j|_{u_j:=u_k:=u}$ denotes the function $\tilde{w}_j(u_1, \dots, u_d)$ evaluated at $u_j = u$ and $u_k = u$. It is easy to check that

$$\begin{aligned} \frac{\partial \tilde{w}_j}{\partial u_k} \Big|_{u_j:=u_k:=u} &= - \sum_{A: j \in A, k \notin A} \frac{\xi'_k(u)}{\xi_k(u)} Q_{A \cup k} \Xi_A \theta_A(u) \\ &\quad + \sum_{A: j, k \in A} \xi'_k(u) Q_A \Xi_{A \setminus k} \theta_A(u) \end{aligned}$$

and

$$\frac{\partial \tilde{w}_k}{\partial u_k} \Big|_{u_j:=u_k:=u} = \sum_{A: j \notin A, k \in A} \frac{\xi'_j(u)}{\xi_j(u)} Q_{A \cup j} \Xi_A \theta_A(u)$$

$$\begin{aligned}
& + \sum_{A:j,k \in A} \sum_{i \notin A} \frac{\xi_i(u_i) \xi'_i(u)}{\xi_i^2(u)} Q_{A \cup i} \Xi_A \theta_A(u) \\
& + \sum_{A:j,k \in A} \xi'_k(u) Q_A \Xi_{A \setminus k} \theta_A(u) \\
& + \sum_{A:j,k \in A} Q_A \Xi_A \theta'_A(u).
\end{aligned}$$

Hence

$$\begin{aligned}
\frac{1}{\prod_i \xi_i(u_i)} \left(\frac{\partial \tilde{w}_k}{\partial u_k} - \frac{\partial \tilde{w}_j}{\partial u_k} \right) \Big|_{u_j := u_k := u} &= \sum_{A:j,k \in A} R_A \theta'_A(u) \\
& + \sum_{A:j,k \in A} \sum_{i \notin A} \frac{\xi'_i(u)}{\xi_i^2(u)} R_{A \cup i} \theta_A(u) \\
& + \sum_{A:j \notin A, k \in A} \frac{\xi'_j(u)}{\xi_j^2(u)} R_{A \cup j} \theta_A(u) \\
& + \sum_{A:j \in A, k \notin A} \frac{\xi'_k(u)}{\xi_k^2(u)} R_{A \cup k} \theta_A(u),
\end{aligned}$$

where R_A is an abbreviation for the following expression:

$$R_A = \frac{Q_A}{\prod_{i \notin A} \xi_i(u_i)} = \prod_{i \notin A} \left(\frac{1}{\xi_i(u_i)} - \frac{1}{\xi_i(u)} \right).$$

Combining the last three sums, we obtain that

$$\frac{1}{\prod_i \xi_i(u_i)} \left(\frac{\partial \tilde{w}_k}{\partial u_k} - \frac{\partial \tilde{w}_j}{\partial u_k} \right) \Big|_{u_j := u_k := u} = \sum_{A:j,k \in A} R_A \left(\theta'_A(u) + \sum_{i \in A} \frac{\xi'_i(u)}{\xi_i^2(u)} \theta_{A \setminus i}(u) \right).$$

Hence, to guarantee that first derivatives of \tilde{w} are continuous across $\tilde{M}_j \cap \tilde{M}_k$ for all $j \neq k$, it suffices to define θ_A so that

$$(36) \quad \theta'_A = - \sum_{i \in A} \frac{\xi'_i}{\xi_i^2} \theta_{A \setminus i}$$

for all A containing two or more elements. If we let $\bar{u}_A = \min_{j \in A} \bar{u}_j$, then $\tilde{T}_A \bar{u}_A$ lies on one of the “back faces” of \tilde{D} (i.e., one of the components is at its upper bound), and so

$$(37) \quad \theta_A(\bar{u}_A) = 0.$$

Also, if A contains exactly one element, say j , then we see from (28), (32), (33), and the fact that w is to agree with γ_j on the j th coordinate axis, that

$$(38) \quad \theta_j(u) = \frac{\gamma_j(\xi_j(u))}{\xi_j(u)}.$$

Hence, starting with sets A of cardinality two and working upward, each θ_A is uniquely determined by (36) and (37). Performing this recursion, we obtain that

$$\theta_A(u) = \sum_{j \in A} \int_{\mathcal{R}_{A \setminus j}(u)} \left(\prod_{i \in A \setminus j} \frac{\xi'_i(u_i)}{\xi_i^2(u_i)} \right) \theta_j(\max_{i \in A \setminus j} u_i) \prod_{i \in A \setminus j} du_i,$$

where $\mathcal{R}_A(u) = \{(u_i)_{i \in A} : u < u_i \leq \bar{u}_i\}$. Finally, we must check second derivatives. Carefully differentiating, we see that

$$\begin{aligned} \frac{\xi_k^2(u)}{\xi'_k(u)} \frac{1}{\prod_i \xi_i(u_i)} \frac{\partial^2 \tilde{w}_j}{\partial u_j \partial u_k} \Big|_{u_j := u_k := u} &= \sum_{A: j, k \in A} \sum_{i \notin A} \xi_k(u) \frac{\xi'_i(u)}{\xi_i^2(u)} R_{A \cup i} \theta_A(u) \\ &+ \sum_{A: j, k \in A} \xi_k(u) \frac{\xi'_j(u)}{\xi_j(u)} R_A \theta_A(u) \\ &+ \sum_{A: j, k \in A} \xi_k(u) R_A \theta'_A(u) \\ &+ \sum_{A: j \in A, k \notin A} \frac{\xi'_k(u)}{\xi_k(u)} R_{A \cup k} \theta_A(u) \\ &- \sum_{A: j \in A, k \notin A} \sum_{i \notin A, i \neq k} \frac{\xi'_i(u)}{\xi_i^2(u)} R_{A \cup i \cup k} \theta_A(u) \\ &- \sum_{A: j \in A, k \notin A} \frac{\xi'_j(u)}{\xi_j(u)} R_{A \cup k} \theta_A(u) \\ &- \sum_{A: j \in A, k \notin A} R_{A \cup k} \theta'_A(u). \end{aligned}$$

Now substituting (36) into the above formula and reindexing so that the θ 's always are subscripted with an A , we obtain that

$$\begin{aligned} \frac{\xi_j^2(u)}{\xi'_j(u)} \frac{\xi_k^2(u)}{\xi'_k(u)} \frac{1}{\prod_i \xi_i(u_i)} \frac{\partial^2 \tilde{w}_j}{\partial u_j \partial u_k} \Big|_{u_j := u_k := u} &= - \sum_{A: j \notin A, k \in A} \xi_k(u) R_{A \cup j} \theta_A(u) \\ &- \sum_{A: j \in A, k \notin A} \xi_j(u) R_{A \cup k} \theta_A(u) \\ &+ \sum_{A: j, k \in A} \xi_j(u) \xi_k(u) R_A \theta_A(u) \\ &+ \sum_{A: j, k \notin A, |A| \geq 1} R_{A \cup j \cup k} \theta_A(u) \\ &- R_{j \cup k} \theta'_j(u) \frac{\xi_j^2(u)}{\xi'_j(u)}. \end{aligned}$$

Interchanging the roles of j and k , we can write the analogous expression for \tilde{w}_k and then subtract to obtain that

$$\begin{aligned} (39) \quad &\frac{\xi_j^2(u)}{\xi'_j(u)} \frac{\xi_k^2(u)}{\xi'_k(u)} \frac{1}{\prod_i \xi_i(u_i)} \left(\frac{\partial^2 \tilde{w}_j}{\partial u_j \partial u_k} - \frac{\partial^2 \tilde{w}_k}{\partial u_j \partial u_k} \right) \Big|_{u_j := u_k := u} \\ &= R_{j \cup k} \left(\theta'_k(u) \frac{\xi_k^2(u)}{\xi'_k(u)} - \theta'_j(u) \frac{\xi_j^2(u)}{\xi'_j(u)} \right). \end{aligned}$$

Now we are almost done. Recalling (38), we see that $\theta_k(u) = \gamma_k(\xi_k(u))/\xi_k(u)$, and so, using (11) and suppressing the dependent variable u , we obtain that

$$\theta'_k \frac{\xi_k^2}{\xi'_k} = \xi_k \gamma'_k(\xi_k) - \gamma_k(\xi_k) = -\Gamma_k(\xi_k) = -u.$$

Hence both sides of the difference on the right-hand side of (39) are equal to $-u$, and so the difference vanishes.

4. Face data. Let $w(x)$ denote a candidate for the value function $v(x)$ for the face-data problem. As in the previous section, it is convenient to work in the system of coordinates defined by (28). Hence the control region N_i described in Theorem 1.2 becomes $\tilde{N}_i = \{(u_1, \dots, u_d) : u_i = \min_j u_j\}$.

Assuming that Theorem 1.2 correctly describes the optimal control regions, we see that $w(x)$ should be linear in x_i on the set N_i . Hence we can “sweep out” $w(x)$ to the boundary of N_i . Using the u_i coordinates, this sweeping becomes

$$(40) \quad \tilde{w}(u_1, \dots, u_d) = \left(1 - \frac{\xi_i(u_i)}{\xi_i(u_j)}\right) \prod_{k \neq i} \gamma_k(u_k) + \frac{\xi_i(u_i)}{\xi_i(u_j)} \tilde{w}(u_1, \dots, u_d)|_{u_i:=u_j},$$

for $(u_1, \dots, u_d) \in \tilde{N}_{i,j} = \{(u_1, \dots, u_d) \in \tilde{D} : u_i \leq u_j \leq \min_{k \neq i,j} u_k\}$. First, we ensure that \tilde{w} is twice continuously differentiable across the boundary between $\tilde{N}_{i,j}$ and $\tilde{N}_{j,i}$. Let $\tilde{w}_{i,j}$ denote the restriction of \tilde{w} to $\tilde{N}_{i,j}$, so that $\tilde{w}_{i,j}$ is given by the right-hand side in (40). From (40), we see that

$$\left. \frac{\partial \tilde{w}_{i,j}}{\partial u_i} \right|_{u_i:=u_j:=u} = -\frac{\xi'_i(u)}{\xi_i(u)} \gamma_j(u) \prod_{k \neq i,j} \tilde{\gamma}_k(u_k) + \frac{\xi'_i(u)}{\xi_i(u)} \tilde{w}_{i,j}|_{u_i:=u_j:=u}$$

and

$$\left. \frac{\partial \tilde{w}_{j,i}}{\partial u_i} \right|_{u_i:=u_j:=u} = \frac{\xi'_j(u)}{\xi_j(u)} \gamma_i(u) \prod_{k \neq i,j} \tilde{\gamma}_k(u_k) - \frac{\xi'_j(u)}{\xi_j(u)} \tilde{w}_{j,i}|_{u_i:=u_j:=u} + \frac{\partial}{\partial u} \tilde{w}_{j,i}|_{u_i:=u_j:=u},$$

where $\tilde{\gamma}_k$ is defined by $\tilde{\gamma}_k(u_k) = \gamma_k(\xi_k(u_k))$. Hence, for first derivatives to match, we need

$$(41) \quad \begin{aligned} \frac{\partial}{\partial u} \tilde{w}_{j,i}|_{u_i:=u_j:=u} - \left(\frac{\xi'_i(u)}{\xi_i(u)} + \frac{\xi'_j(u)}{\xi_j(u)} \right) \tilde{w}_{j,i}|_{u_i:=u_j:=u} \\ = - \left(\frac{\xi'_i(u)}{\xi_i(u)} \gamma_j(u) + \frac{\xi'_j(u)}{\xi_j(u)} \gamma_i(u) \right) \prod_{k \neq i,j} \tilde{\gamma}_k(u_k). \end{aligned}$$

At this point, let us consider that part of the state space where $u_1 \leq u_2 \leq \dots \leq u_d$. From (40) and (41), we see that

$$(42) \quad \tilde{w}(u_1, \dots, u_d) = \left(1 - \frac{\xi_1(u_1)}{\xi_1(u_2)}\right) G_1 + \frac{\xi_1(u_1)}{\xi_1(u_2)} \theta_2(u_2),$$

where G_i is an abbreviation for $\prod_{k>i} \tilde{\gamma}_k(u_k)$, and $\theta_2(u) = \tilde{w}(u, u, u_3, \dots, u_d)$ is a solution of

$$\theta'_2 - \left(\frac{\xi'_1}{\xi_1} + \frac{\xi'_2}{\xi_2} \right) \theta_2 = -G_2 \left(\frac{\xi'_1}{\xi_1} \gamma_2 + \frac{\xi'_2}{\xi_2} \gamma_1 \right).$$

Using the integrating factor $1/\xi_1 \xi_2$, we can solve for θ_2 as follows:

$$\theta_2(u) = \xi_1(u) \xi_2(u) \left(\frac{\theta_2(u_3)}{\xi_1(u_3) \xi_2(u_3)} + G_2 \int_u^{u_3} \left(\frac{\xi'_1 \tilde{\gamma}_2}{\xi_1^2 \xi_2} + \frac{\xi'_2 \tilde{\gamma}_1}{\xi_2^2 \xi_1} \right) \right),$$

where we suppress the integration variable and its differential in the above integral. To keep notations in check, we suppress integration variables and differentials several times in the following expressions. We hope that this adds to the clarity of the formulas. Substituting this formula for θ_2 into (42), we obtain that

$$(43) \quad \begin{aligned} \tilde{w}(u_1, \dots, u_d) = & \Xi_1 G_1 \int_{u_1}^{u_2} \frac{\xi'_1}{\xi_1^2} \\ & + \Xi_2 G_2 \int_{u_2}^{u_3} \left(\frac{\xi'_1 \tilde{\gamma}_2}{\xi_1^2 \xi_2} + \frac{\xi'_2 \tilde{\gamma}_1}{\xi_2^2 \xi_1} \right) \\ & + \frac{\Xi_2}{\xi_1(u_3) \xi_2(u_3)} \theta_3(u_3), \end{aligned}$$

where Ξ_i is an abbreviation for $\prod_{k \leq i} \xi_k(u_k)$,

$$\theta_3(u) = \tilde{w}(u, u, u, u_4, \dots, u_d),$$

and the first term from (42) has been rewritten using the following simple identity:

$$\left(1 - \frac{\xi_1(u_1)}{\xi_1(u_2)} \right) = \xi_1(u_1) \int_{u_1}^{u_2} \frac{\xi'_1}{\xi_1^2}.$$

Equation (41) can be used to obtain a directional derivative of \tilde{w} at a point of the form $(u, u, u, u_4, \dots, u_d)$. Writing the analogous expressions obtained by considering the cases where $u_2 \geq \max(u_1, u_3)$ and $u_1 \geq \max(u_2, u_3)$, we can obtain two more independent directional derivatives. From these three independent directional derivatives, it is easy to see that θ_3 satisfies the following differential equation:

$$\theta'_3 - \left(\frac{\xi'_1}{\xi_1} + \frac{\xi'_2}{\xi_2} + \frac{\xi'_3}{\xi_3} \right) \theta_3 = - \left(\frac{\xi'_1}{\xi_1} \tilde{\gamma}_2 \tilde{\gamma}_3 + \frac{\xi'_2}{\xi_2} \tilde{\gamma}_1 \tilde{\gamma}_3 + \frac{\xi'_3}{\xi_3} \tilde{\gamma}_1 \tilde{\gamma}_2 \right).$$

The integrating factor for this differential equation is $1/\xi_1 \xi_2 \xi_3$. Integrating to solve for θ_3 and substituting into (43), we obtain that

$$\begin{aligned} \tilde{w}(u_1, \dots, u_d) = & \Xi_1 G_1 \int_{u_1}^{u_2} \frac{\xi'_1}{\xi_1^2} \\ & + \Xi_2 G_2 \int_{u_2}^{u_3} \left(\frac{\xi'_1 \tilde{\gamma}_2}{\xi_1^2 \xi_2} + \frac{\xi'_2 \tilde{\gamma}_1}{\xi_2^2 \xi_1} \right) \\ & + \Xi_3 G_3 \int_{u_3}^{u_4} \left(\frac{\xi'_1 \tilde{\gamma}_2 \tilde{\gamma}_3}{\xi_1^2 \xi_2 \xi_3} + \frac{\xi'_2 \tilde{\gamma}_1 \tilde{\gamma}_3}{\xi_1 \xi_2^2 \xi_3} + \frac{\xi'_3 \tilde{\gamma}_1 \tilde{\gamma}_2}{\xi_1 \xi_2 \xi_3^2} \right) \\ & + \frac{\Xi_3}{\xi_1(u_4) \xi_2(u_4) \xi_3(u_4)} \theta_4(u_4), \end{aligned}$$

where $\theta_4(u) = \tilde{w}(u, u, u, u, u_5, \dots, u_d)$. Now it is easy to see how this process must continue. Ultimately, we obtain that

$$\tilde{w}(u_1, \dots, u_d) = \sum_{i=1}^d \Xi_i G_i \int_{u_i}^{u_{i+1}} \sum_{k=1}^i \frac{\xi'_k}{\xi_k \tilde{\gamma}_k} \prod_{j \leq i} \frac{\tilde{\gamma}_j}{\xi_j},$$

where we put $u_{d+1} = \bar{u}$ and use the fact that \tilde{w} vanishes at $(\bar{u}, \dots, \bar{u})$. On the other parts of the state space, it is clear that we obtain an analogous formula with the

indices changed so that an index i is replaced with the index of the i th smallest u value.

Now \tilde{w} is defined everywhere in \tilde{D} and is continuously differentiable throughout. It remains to show that it is also twice continuously differentiable. It is sufficient to show that second derivatives agree across the boundary between $\tilde{N}_{i,j}$ and $\tilde{N}_{j,i}$. Straightforward calculation shows that

$$(44) \quad \frac{\xi_i \xi_j}{\xi'_i \xi'_j} \left(\frac{\partial^2 \tilde{w}_{i,j}}{\partial u_i \partial u_j} - \frac{\partial^2 \tilde{w}_{j,i}}{\partial u_i \partial u_j} \right) \Big|_{u_i := u_j := u} = \left(\frac{\xi_i(u)}{\xi'_i(u)} \tilde{\gamma}'_i(u) - \tilde{\gamma}_i(u) - \frac{\xi_j(u)}{\xi'_j(u)} \tilde{\gamma}'_j(u) + \tilde{\gamma}_j(u) \right) \prod_{k \neq i,j} \tilde{\gamma}_k.$$

Since $\tilde{\gamma}_i(u) = \gamma_i(\xi_i(u))$, we see that, by (11),

$$\frac{\xi_i(u)}{\xi'_i(u)} \tilde{\gamma}'_i(u) - \tilde{\gamma}_i(u) = \xi_i(u) \gamma'_i(\xi_i(u)) - \gamma_i(\xi_i(u)) = -\Gamma_i(\xi_i(u)) = -u.$$

Similarly, the last two terms in parentheses in (44) together equal u , and so the right-hand side vanishes. This completes the proof that w is twice continuously differentiable in D .

5. Other examples. Perhaps the most natural extension of the preceding results would be to consider more general biconcave data on the d faces adjacent to the origin. It seems that such generalizations are quite difficult. Indeed, for $d = 3$, we consider the following data:

$$f(x) = \begin{cases} \gamma_0(x_2)\gamma_1(x_3) & \text{if } x = (0, x_2, x_3), \\ \gamma_0(x_3)\gamma_1(x_1) & \text{if } x = (x_1, 0, x_3), \\ \gamma_0(x_1)\gamma_1(x_2) & \text{if } x = (x_1, x_2, 0), \\ 0 & \text{otherwise,} \end{cases}$$

where γ_0 and γ_1 are two different strongly concave functions that vanish at the end-points of their domains. By considering a discretization of the Dirichlet problem (17), (18), we can apply the method of successive approximations to numerically solve for the value function and hence the optimal strategy. As in Fig. 2, there are switching surfaces emanating from the three coordinate axes, but this time they do not meet along a single curve. In fact, the behavior of the optimal switching strategy is quite intricate inside a triangular tube enclosed by the three surfaces. It would be very interesting to understand more about the nature of examples such as this one.

REFERENCES

- [1] R. DALANG, *Randomization in the two-armed bandit problem*, Ann. Probab., 18 (1990), pp. 218–225.
- [2] I. KARATZAS, *Gittins indices in the dynamic allocation problem for diffusion processes*, Ann. Probab., 12 (1984), pp. 173–192.
- [3] A. MANDELBAUM, *Continuous multi-armed bandits and multi-parameter processes*, Ann. Probab., 15 (1987), pp. 1527–1556.
- [4] A. MANDELBAUM, L. SHEPP, AND R. VANDERBEI, *Optimal switching between a pair of Brownian motions*, Ann. Probab., 18 (1990), pp. 1010–1033.
- [5] J. WALSH, *Optional increasing paths*, in Colloque ENST-CNET, Lecture Notes in Math 863, Springer-Verlag, Berlin, New York, 1981, pp. 172–201.
- [6] P. WHITTLE, *Optimization over Time: Dynamic Programming and Stochastic Control*, John Wiley, New York, 1982.

APPROXIMATION IN CONTROL OF THERMOELASTIC SYSTEMS*

J. S. GIBSON[†], I. G. ROSEN[‡], AND G. TAO[§]

Abstract. This paper develops an abstract framework for analysis and approximation of linear thermoelastic control systems, and for design of finite-dimensional compensators. The thermoelastic systems in this paper consist of abstract wave and diffusion equations coupled in a skew self-adjoint fashion. Linear semigroup theory is used to establish that the abstract thermoelastic models are well posed and to prove convergence of generic approximation schemes. Open-loop uniform exponential stability for a subclass of thermoelastic systems is proved via a Lyapunov function. An example involving the design of an optimal linear-quadratic-Gaussian (LQG) compensator for a thermoelastic rod illustrates the application of the abstract theory. Results of an extensive numerical study, including a comparison of the closed-loop performance of different compensator designs, are presented and discussed.

Key words. control theory, optimal control, thermoelastic, approximation, stability

AMS(MOS) subject classifications. 35M05, 65J10, 73C25, 93C25, 49A22

1. Introduction. The transfer of energy between its mechanical form and heat generally has been ignored as a source of both structural damping and excitation in the vast literature on control of flexible structures. Only a few recent papers have considered control of thermoelastic structures [4], [5], [23], [21], [22], [18], [16], [17], [29]. However, the thermally induced vibrations that hampered the recently launched Hubble space telescope have highlighted the coupling between mechanical vibration and heat transfer and the need to model and control thermoelastic phenomena in flexible structures.

This paper has two main objectives: first, to develop a theoretical framework for analysis and approximation in the design of feedback control systems for a broad class of linear thermoelastic systems; second, to illustrate the application of the theory by presenting the most interesting results from an extensive numerical study of linear-quadratic-Gaussian (LQG) optimal control of a thermoelastic rod. Both the theory and the example focus on numerical methods and convergence analysis for the design of finite-dimensional compensators based on finite-dimensional approximations of distributed models of thermoelastic systems.

By a thermoelastic system, we mean an abstract wave equation coupled in a skew self-adjoint fashion with a diffusion equation. While some of the theory developed here pertains specifically to problems in which the generalized wave equation is second-order (in time), much of the theory applies to a broader class of problems, including, for example, problems in which a Schrodinger equation is coupled with a diffusion equation. In this paper, we are particularly interested in second-order generalized wave equations because they are common in flexible structures, but the results here that allow a more general class of wave equations are intended to apply also to problems such as thermal blooming in lasers [34]. Although the theoretical framework developed in this paper handles a wide variety of thermoelastic systems,

* Received by the editors July 16, 1990; accepted for publication (in revised form) June 17, 1991.

[†] Mechanical, Aerospace and Nuclear Engineering, University of California, Los Angeles, California 90024. The work of this author was supported by Air Force Office of Scientific Research grant 87-0373.

[‡] Center for Applied Mathematical Sciences, Department of Mathematics, University of Southern California, Los Angeles, California 90089. The work of this author was supported by Air Force Office of Scientific Research grant 87-0356.

[§] Department of Electrical Engineering, University of Virginia, Charlottesville, Virginia 22903.

it is not clear whether our hypotheses hold for the thermo-viscoelastic systems with memory studied by Burns et al. [4], [5] and Liu [23].

Our philosophy in the abstract formulation of thermoelastic control systems in §2 and in the approximation theory in §4 is to base the results on hypotheses that require as little as possible beyond conditions that normally hold for the individual wave and diffusion equations. This means that, in analyzing a particular application, most of the work is done on the uncoupled wave and diffusion equations, and the work required to couple the systems is minimized. For example, in verifying the hypotheses for Theorem 4.6, which concerns convergence of approximations to the open-loop thermoelastic system, once the convergence conditions for independent approximations to the uncoupled wave and diffusion equations are verified, no further work is necessary to guarantee convergence of the approximations to the thermoelastic system when the straightforward Galerkin scheme that we assume for approximating the coupling operator is used.

The approach to compensator design in §3 and 4.1 of this paper is to approximate an ideal infinite-dimensional LQG compensator with a sequence of finite-dimensional compensators. However, the abstract formulation of thermoelastic control systems in §2, the approximation and convergence theory in §4.2, and the result in §5 on open-loop uniform exponential stability should be useful in any method for analysis and design of controllers for thermoelastic systems.

An important issue in both convergence of the approximating compensators and performance of the closed-loop systems is uniform exponential stability of the open-loop thermoelastic system. While several authors [18], [9], [21], [31], [33] have proved strong stability for various linear and nonlinear thermoelastic systems, few results have been published on uniform exponential stability. A result in [31] on integrability of the energy, when applied to the linear case, yields uniform exponential stability for thermoelastic rods with certain sets of boundary conditions. Also, a recent eigenvalue analysis in [18] yields uniform exponential stability for linear thermoelastic rods with the same sets of boundary conditions to which the result in [31] applies. The proof of our Theorem 5.1 uses a Lyapunov function to establish uniform exponential stability for a large class of linear thermoelastic systems, but does not improve on the results in [18] and [31] for the rod. The results in [18], [31] and our §5 do not apply to the case of a linear thermoelastic rod with all Dirichlet boundary conditions, for which uniform exponential stability has been proven recently in [20], [24], [25], [6].

In §6, we apply the theory developed in §§2–5 to design finite-dimensional compensators for a thermoelastic rod. We present numerical results for the functional control and estimator gains that represent the compensators graphically. We also compare the closed-loop eigenvalues produced by three of the finite-dimensional compensators based on different damping models. These eigenvalues were obtained from simulations in which each compensator was connected to a model of the rod with dimension significantly higher than the dimension of the compensator. This comparison illustrates the importance of modelling even very light thermoelastic damping, or possibly an artificial viscous equivalent, if no stronger damping mechanism is present.

2. Abstract thermoelastic systems. Throughout this paper, H or H_j ($j = 0, 1, 2$) will be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$ or $\langle \cdot, \cdot \rangle_j$ and corresponding induced norm $|\cdot|$ or $|\cdot|_j$. Also, V or V_j will be a reflexive Banach space with norm $\|\cdot\|$ or $\|\cdot\|_j$. The continuous dual of V will be denoted by V' , and

$$(2.1) \quad V \hookrightarrow H \hookrightarrow V'$$

will mean that V is embedded densely and continuously in H , which implies that H is embedded densely and continuously in V' (see, for example, [30], [32]). In this case, $\langle \cdot, \cdot \rangle$ will denote both the H -inner product and the duality pairing on $V \times V'$.

LEMMA 2.1. *Let V and H be related as in (2.1), let \mathcal{A} be a linear isomorphism (i.e., a continuous linear bijection with continuous inverse) from V to V' such that \mathcal{A} is dissipative in the sense that*

$$(2.2) \quad \operatorname{Re} \langle v, \mathcal{A}v \rangle \leq 0 \quad \forall v \in V,$$

and define

$$(2.3) \quad \operatorname{Dom}(A) = \mathcal{A}^{-1}H, \quad A = \mathcal{A}|_{\operatorname{Dom}(A)}.$$

Then $\operatorname{Dom}(A)$ is dense in H and $A^{-1} \in \mathcal{B}(H, H)$. Also, A is a maximal dissipative operator on H .

Proof. That $\operatorname{Dom}(A)$ is dense in H follows from the fact that H is dense in V' and \mathcal{A}^{-1} is bounded from V' to H . To see that A is maximal dissipative, suppose that there exists a dissipative linear operator $\tilde{A} : \operatorname{Dom}(\tilde{A}) \subset H \rightarrow H$ that is a proper extension of A . Since $\mathcal{R}(A) = H$, there exists $h \in \operatorname{Dom}(\tilde{A}) \setminus \operatorname{Dom}(A)$ and $v \in \operatorname{Dom}(A)$ such that $h \neq 0$, $\tilde{A}h = 0$, and $Av = h$. Then, for any real α , $\langle v + \alpha h, \tilde{A}(v + \alpha h) \rangle = \langle v, Av \rangle + \alpha |h|^2$, and, for sufficiently large $\alpha > 0$, $\operatorname{Re} \langle v + \alpha h, \tilde{A}(v + \alpha h) \rangle > 0$, contradicting the dissipativity of \tilde{A} . \square

THEOREM 2.2. *Let the Hilbert space H_1 , the reflexive Banach space V_1 , and the operator \mathcal{A}_1 be as in Lemma 2.1. Let the Hilbert space H_2 and the reflexive Banach space V_2 be as in Lemma 2.1, and let \mathcal{A}_2 be a linear isomorphism from V_2 to V'_2 that is V_2 -coercive; i.e., there exists a positive real number α such that*

$$(2.4) \quad \operatorname{Re} \langle \phi, \mathcal{A}_2 \phi \rangle_2 \geq \alpha \|\phi\|_2^2, \quad \phi \in V_2.$$

Also, let $\mathcal{L} \in \mathcal{B}(V_1, V'_2)$. Define

$$(2.5) \quad H = H_1 \times H_2, \quad V = V_1 \times V_2$$

and

$$(2.6) \quad \mathcal{A} = \begin{bmatrix} \mathcal{A}_1 & -\mathcal{L}^* \\ \mathcal{L} & -\mathcal{A}_2 \end{bmatrix}$$

where $\mathcal{L}^* \in \mathcal{B}(V_2, V'_1)$ is defined by

$$(2.7) \quad \langle \psi, \mathcal{L}^* \phi \rangle_1 = \overline{\langle \phi, \mathcal{L} \psi \rangle_2}, \quad \psi \in V_1, \quad \phi \in V_2$$

(i.e., \mathcal{L}^* is the Banach-space adjoint of \mathcal{L}). Then H , V , V' and \mathcal{A} are as in Lemma 2.1.

Proof. Since \mathcal{A}_1 is dissipative and \mathcal{A}_2 is V_2 -coercive, the operator $(\mathcal{A}_2 - \mathcal{L}\mathcal{A}_1^{-1}\mathcal{L}^*) \in \mathcal{B}(V_2, V'_2)$ is V_2 -coercive. Hence, for $f_1 \in V'_1$ and $f_2 \in V'_2$, the pair $(v_1, v_2) \in V$ given by

$$(2.8) \quad v_2 = (\mathcal{A}_2 - \mathcal{L}\mathcal{A}_1^{-1}\mathcal{L}^*)^{-1}(\mathcal{L}\mathcal{A}_1^{-1}f_1 - f_2), \quad v_1 = \mathcal{A}_1^{-1}(\mathcal{L}^*v_2 + f_1)$$

is the unique solution to

$$(2.9) \quad \mathcal{A} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix}.$$

The mapping that takes (f_1, f_2) to (v_1, v_2) is clearly bounded from $V' = V'_1 \times V'_2$ to V . \square

Remark 2.3. We define the adjoint operators $\mathcal{A}_1^* \in \mathcal{B}(V_1, V'_1)$, $\mathcal{A}_2^* \in \mathcal{B}(V_2, V'_2)$, and $\mathcal{A}^* \in \mathcal{B}(V, V')$ as in (2.7) with the appropriate duality pairing in each case. Under the hypotheses of Theorem 2.2, \mathcal{A}_1^* , \mathcal{A}_2^* , and \mathcal{A}^* have the same properties, respectively, as \mathcal{A}_1 , \mathcal{A}_2 , and \mathcal{A} .

Remark 2.4. We define the operator $L : \text{Dom}(L) \subset H_1 \rightarrow H_2$ to be the restriction of \mathcal{L} to $\text{Dom}(L) = \{\psi \in V_1 : \mathcal{L}\psi \in H_2\}$. If $\text{Dom}(L)$ is dense in V_1 , we define the operator $L^* : \text{Dom}(L^*) \subset H_2 \rightarrow H_1$ to be the Hilbert space adjoint of L with respect to the H_1 and H_2 inner products. It can be shown that L^* is the restriction of \mathcal{L}^* to $\text{Dom}(L^*) = \{\phi \in V_2 : \mathcal{L}^*\phi \in H_1\}$.

For the class of systems of primary interest in this paper, there exist Hilbert spaces H_0 and H_2 and reflexive Banach spaces V_0 and V_2 such that $V_0 \hookrightarrow H_0 \hookrightarrow V'_0$ and $V_2 \hookrightarrow H_2 \hookrightarrow V'_2$ (with each injection continuous and dense). The thermoelastic evolution equations have the form

$$(2.10) \quad \ddot{w}(t) + \mathcal{D}_0 \dot{w}(t) + \mathcal{A}_0 w(t) + \mathcal{L}_0^* \theta(t) = f_0(t), \quad t > 0,$$

$$(2.11) \quad \dot{\theta}(t) + \mathcal{A}_2 \theta(t) - \mathcal{L}_0 \dot{w}(t) = f_2(t), \quad t > 0,$$

where $\mathcal{D}_0, \mathcal{A}_0 \in \mathcal{B}(V_0, V'_0)$, $\mathcal{L}_0 \in \mathcal{B}(V_0, V'_2)$, $\mathcal{A}_2 \in \mathcal{B}(V_2, V'_2)$, $f_i \in L_1(0, \bar{t}; H_i)$ for $i = 0, 2$ and all $\bar{t} > 0$.

We assume that \mathcal{A}_0 is symmetric in the sense that

$$(2.12) \quad \langle \psi, \mathcal{A}_0 \phi \rangle_0 = \overline{\langle \phi, \mathcal{A}_0 \psi \rangle_0}, \quad \phi, \psi \in V_0,$$

and that \mathcal{A}_0 is V_0 -coercive and \mathcal{A}_2 is V_2 -coercive. We assume that \mathcal{D}_0 is nonnegative in the sense that

$$(2.13) \quad \text{Re} \langle \psi, \mathcal{D}_0 \psi \rangle_0 \geq 0, \quad \psi \in V_0.$$

To derive a semigroup generator for the thermoelastic system in (2.10) and (2.11), we first consider the semigroup generator corresponding to (2.10) for the case $\mathcal{L}_0 = 0$. We make V_0 into a Hilbert space by defining

$$(2.14) \quad \langle \psi, \phi \rangle_{V_0} = \langle \psi, \mathcal{A}_0 \phi \rangle_0, \quad \phi, \psi \in V_0.$$

Our hypotheses on \mathcal{A}_0 imply that the norm induced by the inner product in (2.14) is equivalent to the original V_0 norm. We define

$$(2.15) \quad H_1 = V_0 \times H_0, \quad V_1 = V_0 \times V_0,$$

and we identify V_0 with V'_0 in the first component of H_1 and V_1 and write $V'_1 = V_0 \times V'_0$. It follows that $V_1 \hookrightarrow H_1 \hookrightarrow V'_1$.

Next we define

$$(2.16) \quad \mathcal{A}_1 = \begin{bmatrix} 0 & I \\ -\mathcal{A}_0 & -\mathcal{D}_0 \end{bmatrix} \in \mathcal{B}(V_1, V'_1).$$

That \mathcal{A}_1 is an isomorphism from V_1 to V'_1 follows from

$$(2.17) \quad \mathcal{A}_1^{-1} = \begin{bmatrix} -\mathcal{A}_0^{-1} \mathcal{D}_0 & -\mathcal{A}_0^{-1} \\ I & 0 \end{bmatrix} \in \mathcal{B}(V'_1, V_1).$$

We define A_1 by (2.3) with A , \mathcal{A} , and H replaced by A_1 , \mathcal{A}_1 , and H_1 , respectively. According to Lemma 2.1, A_1 generates a contraction semigroup on H_1 . (See [30], [32], [2], [13] for similar approaches to obtaining semigroup generators of the form in (2.16).) Also, we note that the restriction of $-A_2$ to $\mathcal{A}_2^{-1}H_2$ generates a uniformly exponentially stable analytic contraction semigroup on H_2 . For the thermoelastic system, we define

$$(2.18) \quad \mathcal{L} = [0 \quad \mathcal{L}_0] \in \mathcal{B}(V_1, V'_2)$$

to obtain the situation in Theorem 2.2 with \mathcal{A}_1 defined by (2.16). The corresponding \mathcal{A} defined by (2.6) is

$$(2.19) \quad \mathcal{A} = \begin{bmatrix} 0 & I & 0 \\ -\mathcal{A}_0 & -\mathcal{D}_0 & -\mathcal{L}_0^* \\ 0 & \mathcal{L}_0 & -\mathcal{A}_2 \end{bmatrix} \in \mathcal{B}(V, V')$$

where

$$(2.20) \quad V = V_0 \times V_0 \times V_2 \hookrightarrow H = V_0 \times H_0 \times H_2 \hookrightarrow V' = V_0 \times V'_0 \times V'_2.$$

The semigroup generator A for the thermoelastic system in (2.10) and (2.11) then is defined by (2.3). Explicitly, the domain of this semigroup generator is

$$(2.21) \quad \text{Dom}(A) = \{(\phi, \psi, \theta) \in V : \mathcal{A}(\phi, \psi, \theta) \in H\}.$$

The system in (2.10) and (2.11) now can be written as

$$(2.22) \quad \dot{x}(t) = Ax(t) + f(t), \quad t > 0,$$

where $x(t) = (w(t), \dot{w}(t), \theta(t)) \in H$ and $f = (0, f_0, f_2) \in L_1(0, \bar{t}; H)$ for all $\bar{t} > 0$. If $\{T(t) : t \geq 0\}$ is the semigroup generated by A , the *mild solution* to the initial value problem consisting of (2.22) and an initial condition $x(0) = (w(0), \dot{w}(0), \theta(0)) \in H$ is

$$(2.23) \quad x(t) = T(t)x(0) + \int_0^t T(t-s)f(s)ds, \quad t \geq 0.$$

3. The LQG optimal control problem. In the abstract thermoelastic system (2.10)–(2.11), we consider inputs of the form

$$(3.1) \quad f(t) = Bu(t) + \tilde{B}\gamma(t), \quad t > 0$$

and an output given by

$$(3.2) \quad y(t) = Cx(t) + \nu(t), \quad t > 0,$$

where x is the mild solution to (2.22), $u(t) \in R^m$, $\gamma(t) \in R^\ell$, $y(t) \in R^p$, $\nu(t) \in R^p$, $B \in \mathcal{B}(R^m, H)$, $\tilde{B} \in \mathcal{B}(R^\ell, H)$, and $C \in \mathcal{B}(H, R^p)$. Also, γ and ν are stationary zero-mean Gaussian white noise processes with covariance matrices Γ and \hat{R} , respectively, and \hat{R} is positive definite.

The linear-quadratic-Gaussian (LQG) optimal control problem is: given the output y in (3.2), choose u to minimize

$$(3.3) \quad J(u) = \lim_{t_f \rightarrow \infty} E \left\{ \frac{1}{t_f} \int_0^{t_f} [\langle Qx(t), x(t) \rangle + u(t)^T Ru(t)] dt \right\}$$

where $Q \in \mathcal{B}(H, H)$ and $R \in R^{m \times m}$ are self-adjoint with Q nonnegative and R positive definite; as in (3.2), x is the mild solution to the thermoelastic system (2.10)–(2.11) (or, equivalently, (2.22)) for the input of the form (3.1).

In view of (2.10) and (2.11), the operator B has the form

$$(3.4) \quad B = \begin{bmatrix} 0 \\ B_0 \\ B_2 \end{bmatrix}$$

where

$$(3.5) \quad B_i = [b_{i1} b_{i2} \cdots b_{im}], \quad b_{ij} \in H_i, \quad j = 1, 2, \dots, m, \quad i = 0, 2.$$

The operator \tilde{B} has the same form. The operator C in (3.2) has the form

$$(3.6) \quad C = [C_{01} \ C_{02} \ C_2],$$

where $C_{01} \in \mathcal{B}(V_0, R^p)$, $C_{02} \in \mathcal{B}(H_0, R^p)$, and $C_2 \in \mathcal{B}(H_2, R^p)$.

Theory for the infinite-dimensional LQG optimal control problem with bounded input and output operators can be found in [1], [8], [11], [14], [13]. We briefly summarize the relevant results and essential features of the theory here. As in finite dimensions, the LQG problem separates into a deterministic linear-quadratic regulator problem on the infinite interval and a dual state estimator, or filtering, problem.

First we consider the regulator problem, which is to choose the control u to minimize the integral in (3.3) when both noise processes in (3.1) and (3.2) are zero, the output operator C is the identity, and $t_f = \infty$. If the operator pair (A, B) is uniformly exponentially stabilizable (i.e., there exists a bounded linear operator K such that $A - BK$ generates a uniformly exponentially stable semigroup on H) and the pair (Q, A) is uniformly exponentially detectable (i.e., the pair (A^*, Q) is uniformly exponentially stabilizable), then there exists a unique nonnegative self-adjoint solution $\Pi \in \mathcal{B}(H, H)$ to the operator algebraic Riccati equation

$$(3.7) \quad A^* \Pi + \Pi A - \Pi B R^{-1} B^* \Pi + Q = 0,$$

with $\Pi(\text{Dom}(A)) \subset \text{Dom}(A^*)$. The optimal control for the infinite-time linear-quadratic regulator problem has the feedback form

$$(3.8) \quad u(t) = -Kx(t), \quad t \geq 0,$$

where

$$(3.9) \quad K = R^{-1} B^* \Pi \in \mathcal{B}(H, R^m).$$

For the filtering problem, we define

$$(3.10) \quad \hat{Q} = \tilde{B} \Gamma \tilde{B}^*.$$

If the pair (C, A) is uniformly exponentially detectable and the pair (A, \hat{Q}) is uniformly exponentially stabilizable, the operator algebraic Riccati equation

$$(3.11) \quad A \hat{\Pi} + \hat{\Pi} A^* - \hat{\Pi} C^* \hat{R}^{-1} C \hat{\Pi} + \hat{Q} = 0$$

admits a unique nonnegative self-adjoint solution $\hat{\Pi} \in \mathcal{B}(H, H)$ with $\hat{\Pi}(\text{Dom}(A^*)) \subset \text{Dom}(A)$. The minimum-variance estimate of $x(t)$ given $y(\tau)$ ($\tau \leq t$) is a mild solution $\hat{x}(t)$ to the evolution equation

$$(3.12) \quad \dot{\hat{x}}(t) = A\hat{x}(t) + Bu(t) + \hat{K}\{y(t) - C\hat{x}(t)\}$$

where

$$(3.13) \quad \hat{K} = \hat{\Pi}C^*\hat{R}^{-1} \in \mathcal{B}(R^p, H).$$

The optimal LQG compensator consists of the filter, or state estimator, in (3.12) and the control law

$$(3.14) \quad u(t) = -K\hat{x}(t), \quad t \geq 0,$$

with the control and filter gain operators given by (3.9) and (3.13), respectively.

The optimal closed-loop system then takes the form

$$(3.15) \quad z(t) = S_{cl}(t-s)z(s), \quad 0 \leq s \leq t$$

where $z(t) = (x(t), \hat{x}(t)) \in Z = H \times H$ and $\{S_{cl}(t) : t \geq 0\}$ is the C_0 -semigroup of bounded linear operators on Z with infinitesimal generator

$$(3.16) \quad A_{cl} = \begin{bmatrix} A & -BK \\ \hat{K}C & A - BK - \hat{K}C \end{bmatrix}, \quad \text{Dom}(A_{cl}) = \text{Dom}(A) \times \text{Dom}(A).$$

If $\{S(t) : t \geq 0\}$ and $\{\hat{S}(t) : t \geq 0\}$ are the semigroups of bounded linear operators generated on H by infinitesimal generators $A - BK$ and $A - \hat{K}C$, respectively, then it is easy to show that

$$(3.17) \quad e(t) = \hat{S}(t)e(0), \quad t \geq 0,$$

where $e(t) = x(t) - \hat{x}(t)$. Moreover, if for some real a and M ,

$$(3.18) \quad \|S(t)\| \leq Me^{-at}, \quad t \geq 0,$$

$$(3.19) \quad \|\hat{S}(t)\| \leq Me^{-at}, \quad t \geq 0,$$

then for each $b < a$, there exists a constant $M_{cl} > 0$ for which

$$(3.20) \quad \|S_{cl}(t)\| \leq M_{cl}e^{-bt}, \quad t \geq 0.$$

Finally, as in the finite-dimensional case, it can be shown that

$$(3.21) \quad \sigma(A_{cl}) = \sigma(A - BK) \cup \sigma(A - \hat{K}C)$$

where $\sigma(A_{cl})$ denotes the spectrum of the closed-loop semigroup generator in (3.16).

We note that the uniform exponential stabilizability and detectability conditions stated in this section are sufficient for the existence of unique nonnegative self-adjoint solutions to the operator algebraic Riccati equations (3.7) and (3.11). These conditions are not necessary for some problems with finite rank Q and \hat{Q} . A sufficient and usually necessary condition for uniform exponential stabilizability and detectability is that the open-loop system be uniformly exponentially stable, except possibly on a controllable and observable finite-dimensional subspace.

It is convenient to note that, since $\mathcal{R}(K) \subset R^m$ and $\text{Dom}(\hat{K}) = R^p$, there exist $k = (k_1, \dots, k_m)$ and $\hat{k} = (\hat{k}_1, \dots, \hat{k}_p)$ with k_j and \hat{k}_j in H such that

$$(3.22) \quad [Kx]_j = \langle x, k_j \rangle, \quad x \in H, \quad j = 1, 2, \dots, m,$$

and

$$(3.23) \quad \hat{K}r = \sum_{j=1}^p \hat{k}_j r_j = [\hat{k}_1 \hat{k}_2 \dots \hat{k}_p]r, \quad r \in R^p.$$

Also, $k_j, \hat{k}_j \in H$ implies that $k_j = (k_{j,1}, k_{j,2}, k_{j,3})$ and $\hat{k}_j = (\hat{k}_{j,1}, \hat{k}_{j,2}, \hat{k}_{j,3})$ with $k_{j,1}, \hat{k}_{j,1} \in V_0$, $k_{j,2}, \hat{k}_{j,2} \in H_0$, and $k_{j,3}, \hat{k}_{j,3} \in H_2$. It follows that

$$(3.24) \quad \langle x, k_j \rangle = \langle \phi, \mathcal{A}_0 k_{j,1} \rangle_0 + \langle \psi, k_{j,2} \rangle_0 + \langle \theta, k_{j,3} \rangle_2$$

for $x = (\phi, \psi, \theta) \in H$. The vectors k_j and \hat{k}_j and their components, $k_{j,i}$ and $\hat{k}_{j,i}$, are referred to as *functional control and estimator (or observer) gains*, respectively.

4. Approximation and convergence.

4.1. Approximation theory for the LQG control problem. An approximation and convergence theory for the optimal LQG problem for infinite-dimensional systems was developed in [11], [3], [15], [13]. Here, we will first briefly summarize the generic theory and then take a closer look at it in the context of abstract thermoelastic control systems.

Hypothesis 4.1. There exists a sequence of finite-dimensional subspaces H^n ($n = 1, 2, \dots$) of H , and sequences of operators $A^n \in \mathcal{B}(H^n, H^n)$, $B^n \in \mathcal{B}(R^m, H^n)$, $\tilde{B}^n \in \mathcal{B}(R^\ell, H^n)$, $Q^n \in \mathcal{B}(H^n, H^n)$, $C^n \in \mathcal{B}(H^n, R^p)$. The operators Q^n are nonnegative and self-adjoint for each n .

From here on, we take $\hat{Q}^n = \tilde{B}^n \Gamma (\tilde{B}^n)^* \in \mathcal{B}(H^n)$.

Hypothesis 4.2. The finite-dimensional algebraic Riccati equations

$$(4.1) \quad (A^n)^* \Pi^n + \Pi^n A^n - \Pi^n B^n R^{-1} (B^n)^* \Pi^n + Q^n = 0$$

and

$$(4.2) \quad A^n \hat{\Pi}^n + \hat{\Pi}^n (A^n)^* - \hat{\Pi}^n (C^n)^* \hat{R}^{-1} C^n \hat{\Pi}^n + \hat{Q}^n = 0$$

admit unique nonnegative self-adjoint solutions $\Pi^n \in \mathcal{B}(H^n, H^n)$ and $\hat{\Pi}^n \in \mathcal{B}(H^n, H^n)$, respectively.

We define gain operators

$$(4.3) \quad K^n = R^{-1} (B^n)^* \Pi^n \in \mathcal{B}(H^n, R^m),$$

and

$$(4.4) \quad \hat{K}^n = \hat{\Pi}^n (C^n)^* \hat{R}^{-1} \in \mathcal{B}(R^p, H^n),$$

for a sequence of finite-dimensional compensators for the control system (2.10)–(2.11) with input of the form (3.1) and output of the form (3.2). The n th compensator is given by

$$(4.5) \quad u^n(t) = -K^n \hat{x}^n(t),$$

$$(4.6) \quad \dot{\hat{x}}^n(t) = A^n \hat{x}^n(t) + B^n u(t) + \hat{K}^n [y(t) - C^n \hat{x}^n(t)].$$

The resulting closed-loop system is then given by

$$(4.7) \quad z^n(t) = S_{cl}^n(t-s)z(s), \quad 0 \leq s \leq t < \infty$$

where $z^n(t) = (x^n(t), \hat{x}^n(t)) \in Z^n \equiv H \times H^n$, and $\{S_{cl}^n(t) : t \geq 0\}$ is the C_0 -semigroup of bounded linear operators on Z^n with infinitesimal generator $A_{cl}^n : \text{Dom}(A_{cl}^n) \subset Z^n \rightarrow Z^n$ given by

$$(4.8) \quad A_{cl}^n = \begin{bmatrix} A & -BK^n \\ \hat{K}^n C & [A^n - B^n K^n - \hat{K}^n C^n] \end{bmatrix}, \quad \text{Dom}(A_{cl}^n) = \text{Dom}(A) \times H^n.$$

Since $K^n \in \mathcal{B}(H^n, R^m)$ and $\hat{K}^n \in \mathcal{B}(R^p, H^n)$, we have

$$(4.9) \quad [K^n x^n]_j = \langle k_j^n, x^n \rangle_H, \quad j = 1, 2, \dots, m$$

for $x^n \in H^n$ and

$$(4.10) \quad \hat{K}^n r = \sum_{j=1}^p \hat{k}_j^n r_j = [\hat{k}_1^n \ \hat{k}_2^n \ \dots \ \hat{k}_p^n] r$$

for $r \in R^p$ with $k_i^n, \hat{k}_j^n \in H^n$, $i = 1, 2, \dots, m$, $j = 1, 2, \dots, p$.

The convergence theory can be summarized as follows. We will refer to the following finite-dimensional semigroups:

$$(4.11) \quad T^n(t) = e^{A^n t}, \quad S^n(t) = e^{[A^n - B^n K^n]t}, \quad \hat{S}^n(t) = e^{[A^n - \hat{K}^n C^n]t},$$

and their adjoints $T^n(t)^*$, $S^n(t)^*$, and $\hat{S}^n(t)^*$.

Hypothesis 4.3. For each n , there exists a linear mapping P^n from H onto H^n such that

$$(4.12) \quad \lim_{n \rightarrow \infty} P^n x = x, \quad x \in H.$$

For each $x \in H$ and each $t \geq 0$,

$$(4.13) \quad \lim_{n \rightarrow \infty} T^n(t) P^n x = T(t)x,$$

$$(4.14) \quad \lim_{n \rightarrow \infty} T^n(t)^* P^n x = T(t)^* x,$$

where, in each case, the convergence is uniform in t for t in bounded intervals. Also,

$$(4.15) \quad \lim_{n \rightarrow \infty} B^n u = Bu, \quad u \in R^m,$$

$$(4.16) \quad \lim_{n \rightarrow \infty} Q^n P^n x = Qx, \quad x \in H,$$

and

$$(4.17) \quad \lim_{n \rightarrow \infty} C P^n x = Cx, \quad x \in H.$$

If

$$(4.18) \quad \sup_n \|\Pi^n\| < \infty \quad \text{and} \quad \sup_n \|\hat{\Pi}^n\| < \infty$$

and there exist positive constants M and a , independent of n , for which

$$(4.19) \quad \|S^n(t)\| \leq Me^{-at}, \quad \text{and} \quad \|\hat{S}^n(t)\| \leq Me^{-at}, \quad t \geq 0,$$

then the algebraic Riccati equations (3.7) and (3.11) admit bounded nonnegative self-adjoint solutions Π and $\hat{\Pi}$, and

$$(4.20) \quad \lim_{n \rightarrow \infty} \Pi^n P^n x = \Pi x, \quad x \in H,$$

$$(4.21) \quad \lim_{n \rightarrow \infty} \hat{\Pi}^n P^n x = \hat{\Pi} x, \quad x \in H.$$

Also,

$$(4.22) \quad \lim_{n \rightarrow \infty} S^n(t) P^n x = S(t)x, \quad x \in H,$$

and

$$(4.23) \quad \lim_{n \rightarrow \infty} \hat{S}^n(t) P^n x = \hat{S}(t)x, \quad x \in H,$$

with the convergence uniform in t in bounded t -intervals. If, in addition, the operators Q^n and \hat{Q}^n are coercive and bounded away from 0 uniformly in n , then the uniform boundedness of $\|\Pi^n\|$ and $\|\hat{\Pi}^n\|$ yields the existence of positive constants M and a independent of n for which (4.19) holds.

The easiest way to guarantee (4.18) and (4.19) is to show that there exist positive constants M and a , independent of n , for which

$$(4.24) \quad \|T^n(t)\| \leq Me^{-at}, \quad t \geq 0,$$

although such a uniform decay rate for the approximating open-loop semigroups does not always exist. When (4.18) holds but the semigroups $\{S^n(t) : t \geq 0\}$ and $\{\hat{S}^n(t) : t \geq 0\}$ are not necessarily uniformly exponentially stable, uniformly in n , then bounded nonnegative self-adjoint solutions Π and $\hat{\Pi}$ to (3.7) and (3.11) exist, but Π_n and $\hat{\Pi}_n$ are guaranteed only to converge weakly to Π and $\hat{\Pi}$, respectively, as $n \rightarrow \infty$.

When the strong convergence in (4.20) and (4.21) holds, we obtain

$$(4.25) \quad \lim_{n \rightarrow \infty} \|K^n P^n - K\|_{\mathcal{B}(H, R^m)} = 0,$$

$$(4.26) \quad \lim_{n \rightarrow \infty} \|\hat{K}^n - \hat{K}\|_{\mathcal{B}(R^p, H)} = 0,$$

and therefore

$$(4.27) \quad \lim_{n \rightarrow \infty} k_j^n = k_j, \quad j = 1, 2, \dots, m,$$

and

$$(4.28) \quad \lim_{n \rightarrow \infty} \hat{k}_j^n = \hat{k}_j, \quad j = 1, 2, \dots, p,$$

in H . If we define $P_{cl}^n : Z \rightarrow Z^n$ by

$$(4.29) \quad P_{cl}^n = \begin{bmatrix} I & 0 \\ 0 & P^n \end{bmatrix},$$

then we obtain further that

$$(4.30) \quad \lim_{n \rightarrow \infty} S_{cl}^n(t) P_{cl}^n z = S_{cl}(t) z, \quad z \in Z,$$

uniformly on bounded t -intervals.

4.2. Abstract approximation theory for linear thermoelastic systems.

Now we consider the construction of the approximating finite-dimensional subspaces H^n , the mappings P^n , and the operators A^n , B^n , Q^n , etc. We establish a generic approximation theory for abstract linear thermoelastic systems that includes relatively easily verified sufficient conditions for the convergence in Hypothesis 4.3.

We assume the hypotheses of Theorem 2.2.

Hypothesis 4.4. For $j = 1, 2$, and $n = 1, 2, 3, \dots$, H_j^n is a finite-dimensional subspace of V_j and $A_j^n \in \mathcal{B}(H_j^n, H_j^n)$ such that the following conditions hold.

(i) For each $v_j \in V_j$ ($j = 1, 2$), there exists a sequence $v_j^n \in H_j^n$ such that

$$(4.31) \quad v_j^n \xrightarrow{V_j} v_j.$$

(ii) For each n , A_1^n is dissipative; i.e.,

$$(4.32) \quad \operatorname{Re} \langle v, A_1^n v \rangle_1 \leq 0, \quad v \in H_1^n.$$

(iii) For each $f \in V_1'$ and each real $\lambda > 0$,

$$(4.33) \quad (\lambda - A_1^n)^{-1} P_1^{n'} f \xrightarrow{V_1} (\lambda - \mathcal{A}_1)^{-1} f$$

and

$$(4.34) \quad (\lambda - A_1^{n*})^{-1} P_1^{n'} f \xrightarrow{V_1} (\lambda - \mathcal{A}_1^*)^{-1} f,$$

where $P_j^{n'} \in \mathcal{B}(V_j', H_j^n)$ is defined by

$$(4.35) \quad \langle v, \tilde{P}_j^n f \rangle_j = \langle v, f \rangle_j, \quad v \in H_j^n, \quad j = 1, 2.$$

(iv) There exists a positive constant α such that, for all n ,

$$(4.36) \quad \operatorname{Re} \langle v, A_2^n v \rangle_2 \geq \alpha \|v\|_2^2, \quad v \in H_2^n.$$

(v) For each $f \in V_2'$ and each real $\lambda > 0$,

$$(4.37) \quad (\lambda + A_2^n)^{-1} \tilde{P}_2^n f \xrightarrow{V_2} (\lambda + \mathcal{A}_2)^{-1} f$$

and

$$(4.38) \quad (\lambda + A_2^{n*})^{-1} \tilde{P}_2^n f \xrightarrow{V_2} (\lambda + \mathcal{A}_2^*)^{-1} f.$$

Remark 4.5. The operator \tilde{P}_j^n restricts a functional $f \in V'_j$ to H_j^n and identifies $f|_{H_j^n}$ with an element of H_j^n via the Riesz map for H_j^n . If f can be identified with an element of H_j (via the Riesz map for H_j), then $\tilde{P}_j^n f$ is the H_j -projection of f onto H_j^n .

With \tilde{P}_j^n defined by (4.35), we define $L^n \in \mathcal{B}(H_1^n, H_2^n)$ and $L^{n*} \in \mathcal{B}(H_2^n, H_1^n)$ by

$$(4.39) \quad L^n = \tilde{P}_2^n \mathcal{L} |_{H_1^n} \quad \text{or} \quad \langle v_2, L^n v_1 \rangle_2 = \langle v_2, \mathcal{L} v_1 \rangle_2, \quad v_1 \in H_1^n, v_2 \in H_2^n,$$

$$(4.40) \quad L^{n*} = \tilde{P}_1^n \mathcal{L}^* |_{H_2^n} \quad \text{or} \quad \langle v_1, L^{n*} v_2 \rangle_1 = \langle v_1, \mathcal{L}^* v_2 \rangle_1, \quad v_1 \in H_1^n, v_2 \in H_2^n.$$

Hence L^{n*} is the Hilbert-space adjoint of L^n . The operator L^n is a straightforward Galerkin approximation of \mathcal{L} . On the other hand, Hypothesis 4.4 does not require that A_1^n and A_2^n be Galerkin approximations. Next we define

$$(4.41) \quad H^n = H_1^n \times H_2^n$$

and

$$(4.42) \quad A^n = \begin{bmatrix} A_1^n & -L^{n*} \\ L^n & -A_2^n \end{bmatrix} \in \mathcal{B}(H^n, H^n).$$

THEOREM 4.6. For $f_1 \in V'_1$, $f_2 \in V'_2$, and $\lambda > 0$,

$$(4.43) \quad (\lambda - A^n)^{-1} \begin{pmatrix} \tilde{P}_1^n f_1 \\ \tilde{P}_2^n f_2 \end{pmatrix} \xrightarrow{V} (\lambda - \mathcal{A})^{-1} \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \quad \text{as } n \rightarrow \infty$$

and

$$(4.44) \quad (\lambda - A^{n*})^{-1} \begin{pmatrix} \tilde{P}_1^n f_1 \\ \tilde{P}_2^n f_2 \end{pmatrix} \xrightarrow{V} (\lambda - \mathcal{A}^*)^{-1} \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} \quad \text{as } n \rightarrow \infty.$$

Proof. For $f_1 \in V'_1$ and $f_2 \in V'_2$, we set

$$(4.45) \quad v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = (\lambda - \mathcal{A})^{-1} \begin{pmatrix} f_1 \\ f_2 \end{pmatrix},$$

and

$$(4.46) \quad v^n = \begin{pmatrix} v_1^n \\ v_2^n \end{pmatrix} = (\lambda - A^n)^{-1} \begin{pmatrix} \tilde{P}_1^n f_1 \\ \tilde{P}_2^n f_2 \end{pmatrix}.$$

We note that (4.46) is equivalent to

$$(4.47) \quad v_1^n = (\lambda - A_1^n)^{-1} (\tilde{P}_1^n f_1 - L^{n*} v_2^n) = (\lambda - A_1^n)^{-1} \tilde{P}_1^n (f_1 - \mathcal{L}^* v_2^n)$$

and

$$(4.48) \quad v_2^n = (\lambda + A_2^n)^{-1} (\tilde{P}_2^n f_2 + L^n v_1^n) = (\lambda + A_2^n)^{-1} \tilde{P}_2^n (f_2 + \mathcal{L} v_1^n).$$

Substituting (4.47) into (4.48) yields

$$(4.49) \quad [(\lambda + A_2^n) + L^n (\lambda - A_1^n)^{-1} L^{n*}] v_2^n = \tilde{P}_2^n f_2 + L^n (\lambda - A_1^n)^{-1} \tilde{P}_1^n f_1.$$

From (4.32), (4.36), (4.39), and (4.49), we have

$$\begin{aligned}
 \alpha \|v_2^n\|_2^2 &\leq \operatorname{Re}(\langle v_2^n, [(\lambda + A_2^n) + L^n(\lambda - A_1^n)^{-1} L^{n*}] v_2^n \rangle_2) \\
 &= \operatorname{Re}(\langle v_2^n, \tilde{P}_2^n f_2 \rangle_2 + \langle v_2^n, L^n(\lambda - A_1^n)^{-1} \tilde{P}_1^n f_1 \rangle_2) \\
 &= \operatorname{Re}(\langle v_2^n, f_2 \rangle_2 + \langle v_2^n, \mathcal{L}(\lambda - A_1^n)^{-1} \tilde{P}_1^n f_1 \rangle_2) \\
 &\leq \|v_2^n\|_2 (\|f_2\|_{V_2'} + \|\mathcal{L}\| \cdot \|(\lambda - A_1^n)^{-1} \tilde{P}_1^n f_1\|_1).
 \end{aligned}
 \tag{4.50}$$

Since (4.33) implies that $\|(\lambda - A_1^n)^{-1} \tilde{P}_1^n f_1\|_1$ is bounded in n , (4.50) shows that $\|v_2^n\|_2^2$ is bounded in n . Then, it follows from (4.33) and (4.47) that $\|v_1^n\|_1$ is bounded in n .

Next, we note that, for $z = (z_1, z_2) \in H^n$,

$$\begin{aligned}
 \operatorname{Re}\langle z, (\lambda - A^n)z \rangle &= \\
 \operatorname{Re}(\langle z_1, (\lambda - A_1^n)z_1 \rangle_1 + \langle z_2, (\lambda + A_2^n)z_2 \rangle_2) &\geq \alpha \|z_2\|_2^2 + \lambda |z|^2.
 \end{aligned}
 \tag{4.51}$$

We set

$$\tilde{v}^n = \begin{pmatrix} \tilde{v}_1^n \\ \tilde{v}_2^n \end{pmatrix} = \begin{pmatrix} (\lambda - A_1^n)^{-1} \tilde{P}_1^n (f_1 - \mathcal{L}^* v_2) \\ (\lambda + A_2^n)^{-1} \tilde{P}_2^n (f_2 + \mathcal{L} v_1) \end{pmatrix}
 \tag{4.52}$$

and

$$z^n = \begin{pmatrix} z_1^n \\ z_2^n \end{pmatrix} = v^n - \tilde{v}^n.
 \tag{4.53}$$

Then, recalling (4.47) and (4.52) yields

$$\begin{aligned}
 \langle z_1^n, (\lambda - A_1^n)z_1^n \rangle_1 &= \langle z_1^n, (\lambda - A_1^n)v_1^n - (\lambda - A_1^n)\tilde{v}_1^n \rangle_1 \\
 &= -\langle z_1^n, \mathcal{L}^*(v_2^n - v_2) \rangle_1 = -\langle z_1^n, \mathcal{L}^* z_2^n \rangle_1 - \langle z_1^n, \mathcal{L}^*(\tilde{v}_2^n - v_2) \rangle_1
 \end{aligned}
 \tag{4.54}$$

and similarly (4.48) gives

$$\begin{aligned}
 \langle z_2^n, (\lambda + A_2^n)z_2^n \rangle_2 &= \langle z_2^n, (\lambda + A_2^n)v_2^n - (\lambda + A_2^n)\tilde{v}_2^n \rangle_2 \\
 &= \langle z_2^n, \mathcal{L}(v_1^n - v_1) \rangle_2 = \langle z_2^n, \mathcal{L}z_1^n \rangle_2 + \langle z_2^n, \mathcal{L}(\tilde{v}_1^n - v_1) \rangle_2.
 \end{aligned}
 \tag{4.55}$$

Hence

$$\begin{aligned}
 \operatorname{Re}(\langle z_1^n, (\lambda - A_1^n)z_1^n \rangle_1 + \langle z_2^n, (\lambda + A_2^n)z_2^n \rangle_2) \\
 = \operatorname{Re}(-\langle z_1^n, \mathcal{L}^*(\tilde{v}_2^n - v_2) \rangle_1 + \langle z_2^n, \mathcal{L}(\tilde{v}_1^n - v_1) \rangle_2).
 \end{aligned}
 \tag{4.56}$$

In view of (4.51) then,

$$\alpha \|v_2^n - \tilde{v}_2^n\|_2^2 \leq \|z_1^n\|_1 \cdot \|\mathcal{L}\| \cdot \|\tilde{v}_2^n - v_2\|_2 + \|z_2^n\|_2 \cdot \|\mathcal{L}\| \cdot \|\tilde{v}_1^n - v_1\|_1.
 \tag{4.57}$$

According to (2.6), (4.45), (4.52) and conditions (iii) and (v) of Hypothesis 4.4,

$$\lim_{n \rightarrow \infty} \|\tilde{v}_1^n - v_1\|_1 = 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} \|\tilde{v}_2^n - v_2\|_2 = 0.
 \tag{4.58}$$

Hence, $\|\tilde{v}^n\|$ is bounded in n , and we have seen that $\|v^n\|$ is bounded in n . Hence $\|z^n\|$ is bounded in n . Therefore, (4.57) and (4.58) show that v_2^n converges in V_2 to v_2 . Then (4.47), condition (iii) of Hypothesis 4.4, and (4.58) show that v_1^n converges

in V_1 to v_1 , and (4.43) is proven. The proof of (4.44) is the same except that all operators except \tilde{P}_1^n and \tilde{P}_2^n are replaced by their adjoints. \square

Hypothesis 4.4 holds for most common approximation schemes, Galerkin schemes, in particular. The following theorem establishes conditions (iv) and (v) of Hypothesis 4.4 when A_2^n represents a Galerkin approximation of \mathcal{A}_2 .

THEOREM 4.7. *Assume the hypotheses of Theorem 2.2 regarding H_2 , V_2 , and \mathcal{A}_2 , and assume condition (i) of Hypothesis 4.4 for $j = 2$. Define $A_2^n \in \mathcal{B}(H_2^n, H_2^n)$ by*

$$(4.59) \quad A_2^n = \tilde{P}_2^n \mathcal{A}_2|_{H_2^n} \quad \text{or} \quad \langle v, A_2^n w \rangle_2 = \langle v, \mathcal{A}_2 w \rangle_2, \quad v, w \in H_2^n.$$

Then conditions (iv) and (v) of Hypothesis 4.4 hold.

Proof. condition (iv) is immediate. To prove (4.37), let $f \in V_2'$ and set

$$(4.60) \quad v = (\lambda + \mathcal{A}_2)^{-1} f,$$

$$(4.61) \quad v^n = (\lambda + A_2^n)^{-1} \tilde{P}_2^n f.$$

Also, let $\tilde{v}^n \in H_2^n$ such that $\|\tilde{v}^n - v\|_2$ converges to 0. Then

$$(4.62) \quad \alpha \|v^n\|_2^2 \leq |\langle v^n, (\lambda + A_2^n) v^n \rangle_2| = |\langle v^n, \tilde{P}_2^n f \rangle_2| = |\langle v^n, f \rangle_2| \leq \|v^n\|_2 \|f\|_{V_2'}.$$

Hence $\|v^n\|_2$ is bounded in n . Next,

$$(4.63) \quad \begin{aligned} \alpha \|v^n - \tilde{v}^n\|_2^2 &\leq |\langle v^n - \tilde{v}^n, (\lambda + A_2^n)(v^n - \tilde{v}^n) \rangle_2| \\ &= |\langle v^n - \tilde{v}^n, \tilde{P}_2^n f \rangle_2 - \langle v^n - \tilde{v}^n, (\lambda + A_2^n) \tilde{v}^n \rangle_2| = |\langle v^n - \tilde{v}^n, f - (\lambda + \mathcal{A}_2) \tilde{v}^n \rangle_2|. \end{aligned}$$

Since \tilde{v}^n converges in V_2 to v and $\mathcal{A}_2 \in \mathcal{B}(V_2, V_2')$, it follows that $\|\tilde{v}^n\|_2$ is bounded in n and $(\lambda + \mathcal{A}_2) \tilde{v}^n$ converges in V_2' to $(\lambda + \mathcal{A}_2)v = f$. Therefore, (4.63) shows that $\|v^n - \tilde{v}^n\|_2$ converges to 0 as $n \rightarrow \infty$, so that v^n converges in V_2 to v .

The proof is the same when \mathcal{A}_2 and A_2^n are replaced by their adjoints. \square

When \mathcal{A}_1 has the form (2.16) and A_1^n is a Galerkin approximation of \mathcal{A}_1 , condition (iii) of Hypothesis 4.4 can be proved either by arguments similar to the proof of Theorem 4.7 or by projection arguments like those in [13]. Also, see [2].

Usually, the operator P^n in Hypothesis 4.3 is the H -projection onto $H^n = H_1^n \times H_2^n$, so that condition (i) of Hypothesis 4.4 guarantees (4.12). In this case, if $f_j \in H_j$, then $P^n(f_1, f_2) = (\tilde{P}_1^n f_1, \tilde{P}_2^n f_2)$ (recall Remark 4.5). Hence, it follows from Theorem 4.6 and the Trotter–Kato theorem [19] that the approximating open-loop semigroups $T_n(t)$ and $T_n^*(t)$ converge as in Hypothesis 4.3.

Also, when P^n is the H^n -projection, it is most common to define the approximating input, state-weighting, and output operators by

$$(4.64) \quad B^n = P^n B$$

$$(4.65) \quad Q^n = P^n Q|_{H^n},$$

and

$$(4.66) \quad C^n = C|_{H^n},$$

so that (4.15), (4.16), and (4.17) follow from (4.12).

4.3. Matrix representations of approximating operators. We assume now that A_1^n has the form in (2.16), that \mathcal{L} has the form in (2.18), and that H_1 and V_1 have the forms in (2.15). Then H_1^n has the form $H_0^n \times H_0^n$ with $H_0^n \subset V_0$. We assume that, for each n , H_0^n is the span of a finite number of basis vectors $e_{0,i}^n$ and H_2^n is the span of a finite number of basis vectors $e_{2,i}^n$. (The spaces H_0^n and H_2^n may have different dimensions.)

Also, we use Galerkin approximations of both \mathcal{A}_1 and \mathcal{A}_2 . The matrix representation of the operator A^n in (4.42) is then

$$(4.67) \quad \text{matrix representation of } A^n = [A^n] = \begin{bmatrix} 0 & I & 0 \\ -M_0^{n-1} K_0^n & -M_0^{n-1} K_4^n & -M_0^{n-1} K_3^{nT} \\ 0 & M_2^{n-1} K_3^n & -M_2^{n-1} K_2^n \end{bmatrix}$$

where

$$(4.68) \quad \begin{aligned} M_0^n &= [\langle e_{0,i}^n, e_{0,j}^n \rangle_0] & M_2^n &= [\langle e_{2,i}^n, e_{2,j}^n \rangle_2] \\ K_0^n &= [\langle e_{0,i}^n, \mathcal{A}_0 e_{0,j}^n \rangle_0] & K_2^n &= [\langle e_{2,i}^n, \mathcal{A}_2 e_{2,j}^n \rangle_2] \\ K_3^n &= [\langle e_{2,i}^n, \mathcal{L}_0 e_{0,j}^n \rangle_2] & K_4^n &= [\langle e_{0,i}^n, \mathcal{D}_0 e_{0,j}^n \rangle_0]. \end{aligned}$$

The matrix representation of the operator B^n in (3.1) and (3.4) is

$$(4.69) \quad [B^n] = \begin{bmatrix} 0 \\ M_0^{n-1} [\langle e_{0,i}^n, b_{0j} \rangle_0] \\ M_2^{n-1} [\langle e_{2,i}^n, b_{2j} \rangle_2] \end{bmatrix},$$

and the matrix representation of the operator \tilde{B}^n is similar. The matrix representation of the operator C in (3.2) and (3.6) is

$$(4.70) \quad [C^n] = [[C_{01} e_{0,i}^n] \quad [C_{02} e_{0,i}^n] \quad [C_2 e_{2,i}^n]].$$

To discuss the matrix representations of the operators Q^n , \hat{Q}^n , Π^n , and $\hat{\Pi}^n$, it is convenient to define basis vectors

$$(4.71) \quad \tilde{e}_{0,i}^n = (e_{0,i}^n, 0, 0) \quad \tilde{e}_{1,i}^n = (0, e_{0,i}^n, 0) \quad \tilde{e}_{2,i}^n = (0, 0, e_{2,i}^n)$$

and the block-diagonal matrix

$$(4.72) \quad M^n = \text{diag}\{M_0^n, K_0^n, M_2^n\}.$$

The matrix representations of Q^n and \hat{Q}^n are

$$(4.73) \quad [Q^n] = M^{n-1} [\langle \tilde{e}_{i',i}^n, Q \tilde{e}_{j',j}^n \rangle], \quad [\hat{Q}^n] = M^{n-1} [\langle \tilde{e}_{i',i}^n, \hat{Q} \tilde{e}_{j',j}^n \rangle], \quad i', j' = 0, 1, 2.$$

The matrix representations $[\Pi^n]$ and $[\hat{\Pi}^n]$ of Π^n and $\hat{\Pi}^n$, respectively, are determined by solving Riccati matrix equations equivalent to the operator equations (4.1) and (4.2). The form of $[\Pi^n]$ is like that of $[Q^n]$, and in general neither of these matrices is symmetric. Hence, rather than solving the matrix representation of (4.1) directly, it is preferable to premultiply the matrix representation of (4.1) by M^n to obtain a

Riccati matrix equation that can be solved for the symmetric matrix $M^n[\Pi^n]$. Also, instead of solving the matrix representation of (4.2), it is preferable to postmultiply the matrix representation of (4.2) by M^{n-1} to obtain a Riccati matrix equation that can be solved for the symmetric matrix $[\hat{\Pi}^n]M^{n-1}$ (see [13]).

Finally, it follows from (4.3) and (4.4) that the approximating functional control and estimator gains in (4.9) and (4.10) are given by

$$(4.74) \quad [k_1^n \ k_2^n \ \cdots \ k_m^n] = \tilde{e}^n M^{n-1} [\Pi^n] M^n [B^n] R^{-1},$$

$$(4.75) \quad [\hat{k}_1^n \ \hat{k}_2^n \ \cdots \ \hat{k}_p^n] = \tilde{e}^n [\hat{\Pi}^n] M^{n-1} [C^n]^T \hat{R}^{-1},$$

where

$$(4.76) \quad \tilde{e}^n = [[\tilde{e}_{0,i}^n] \ [\tilde{e}_{1,i}^n] \ [\tilde{e}_{1,i}^n]]$$

and $[\tilde{e}_{0,i}^n]$, for example, is the row matrix containing the basis vectors $\tilde{e}_{0,i}^n$ in order. See [13] for details on computing similar functional gains.

5. Stability of the open-loop system. We consider the system in (2.10) and (2.11), and we define

$$(5.1) \quad \text{Dom}(A_0) = \mathcal{A}_0^{-1} H_0, \quad A_0 = \mathcal{A}_0|_{\text{Dom}(A_0)}.$$

Since \mathcal{A}_0 is symmetric and V_0 -coercive, A_0 is self-adjoint and V_0 -coercive. We recall the operators \mathcal{L} and \mathcal{L}_0 in (2.18) and note that $\mathcal{L}_0 \in \mathcal{B}(V_0, V_2')$.

In this section, we assume that

$$(5.2) \quad \mathcal{L}_0 = L_0 \in \mathcal{B}(V_0, H_2),$$

and we assume that there exists a positive real number α such that

$$(5.3) \quad \text{Dom}(A_0) = \{v \in \text{Dom}(L_0) : L_0 v \in \text{Dom}(L_0^*)\} \quad \text{and} \quad A_0 = \alpha L_0^* L_0,$$

where L_0^* is the Hilbert-space adjoint of L_0 with respect to the H_0 and H_2 inner products (recall Remark 2.4). In this case,

$$(5.4) \quad \langle v, w \rangle_{V_0} = \alpha \langle L_0 v, L_0 w \rangle_0, \quad v, w \in V_0.$$

The conditions (5.2) and (5.3) are common in thermoelastic structures because the thermal stress enters the equation governing mechanical vibrations in the same way as the stress due to elastic deformation [7], [27].

THEOREM 5.1. *Assume the conditions stated so far in this section and that the damping operator \mathcal{D}_0 is symmetric (in the sense of (2.12)). If the range of the operator*

$$(5.5) \quad \Lambda_0 = L_0 A_0^{-1}$$

is in V_2 or if \mathcal{D}_0 is H_0 -coercive, then the semigroup generated on the space H in (2.20) by the operator A defined in (2.19)–(2.21) is uniformly exponentially stable.

Proof. First consider the case where $\mathcal{R}(\Lambda_0) \subset V_2$ but \mathcal{D}_0 is not necessarily H_0 -coercive. It is clear that $\Lambda_0 \in \mathcal{B}(H_0, H_2)$ and $\Lambda_0^* \in \mathcal{B}(H_2, H_0)$. Hence, $\mathcal{R}(\Lambda_0) \subset V_2$ implies $\Lambda_0 \in \mathcal{B}(H_0, V_2)$. Furthermore, it can be shown that $\Lambda_0^* \in \mathcal{B}(H_2, V_0)$ and $\alpha L_0 \Lambda_0^*$ is the H_2 -projection onto $\mathcal{R}(L_0)$.

Now define the following self-adjoint bounded linear operator on H :

$$(5.6) \quad Q = \begin{bmatrix} \sigma I & \mathcal{A}_0^{-1} & 0 \\ I & \sigma I & -2\alpha\Lambda_0^* \\ 0 & -2\alpha\Lambda_0 & \sigma I \end{bmatrix}$$

where σ is a positive real number. For σ sufficiently large, Q is H -coercive. Also, since $\mathcal{R}(\Lambda_0) \subset V_2$ and $\mathcal{R}(\Lambda_0^*) \subset V_0$, $QV \subset V$. For $\mathcal{D}_0 = 0$ and $z = (v, h, \theta) \in \text{Dom}(A) \subset V$,

$$(5.7) \quad \begin{aligned} \text{Re} \langle Qz, Az \rangle &= \text{Re} \langle Qz, \mathcal{A}z \rangle = \\ &= -\|v\|_0^2 - \|h\|_0^2 - \sigma \text{Re} \langle \theta, \mathcal{A}_2 \theta \rangle_2 \\ &+ (2\alpha - 1) \text{Re} \langle \theta, L_0 v \rangle_2 + 2\alpha \langle L_0 \Lambda_0^* \theta, \theta \rangle_2 + 2\alpha \text{Re} \langle \Lambda_0 h, \mathcal{A}_2 \theta \rangle_2. \end{aligned}$$

Since

$$(5.8) \quad |\langle \Lambda_0 h, \mathcal{A}_2 \theta \rangle_2| \leq \|\Lambda_0 h\|_2 \cdot \|\mathcal{A}_2\|_{\mathcal{B}(V_2, V_2')} \cdot \|\theta\|_2,$$

and $\Lambda_0 \in \mathcal{B}(H_0, V_2)$, it follows from (5.7) that, for σ sufficiently large, there exists a positive real number β such that

$$(5.9) \quad \text{Re} \langle Qz, Az \rangle \leq -\beta|z|^2, \quad z \in \text{Dom}(A).$$

When $\mathcal{D}_0 \neq 0$, the right side of (5.7) has more terms, but (5.9) can be obtained in a similar manner. The generalized Schwarz inequality $|\langle v, \mathcal{D}_0 h \rangle_0|^2 \leq |\langle v, \mathcal{D}_0 v \rangle_0| \cdot |\langle h, \mathcal{D}_0 h \rangle_0|$ is useful.

If \mathcal{D}_0 is H_0 -coercive, then replacing α with 0 in (5.6) allows (5.9) to be obtained for σ sufficiently large and some positive β . \square

Remark 5.2. The condition $\mathcal{R}(\Lambda_0) \subset V_2$ is equivalent to the following two conditions combined:

$$(5.10) \quad \text{Dom}(L_0^*) \cap \mathcal{N}(L_0^*)^\perp \subset V_2$$

and there exists a real number μ such that

$$(5.11) \quad \|v\|_2 \leq \mu |L_0^* v|_0, \quad v \in \text{Dom}(L_0^*) \cap \mathcal{N}(L_0^*)^\perp.$$

Remark 5.3. To generalize Theorem 5.1 to the case where \mathcal{D}_0 is not symmetric, we would have to impose further conditions on \mathcal{D}_0 , which would take us beyond the focus of this paper.

The hypotheses of Theorem 5.1 hold for many but not all linear thermoelastic systems that seem likely to be uniformly exponentially stable. In most applications, the conditions (5.10) and (5.11) restrict the combinations of boundary conditions. For example, if (2.10) and (2.11) represent a thermoelastic rod, as in the example in the next section, (5.10) and (5.11) hold for Dirichlet boundary conditions on the wave equation at both ends of the rod and Nuemann boundary conditions on the heat equation at both ends, and for various other combinations. However, (5.10) and (5.11) do not hold for Dirichlet boundary conditions on both equations at both ends of the rod.

Recently, it has been proved that the linear thermoelastic rod with all Dirichlet boundary conditions is uniformly exponentially stable [20], [24], [25], [6]. It is interesting that, while [18] showed that all of the eigenvalues are bounded strictly to the

left of the imaginary axis for the linear thermoelastic rod with all Dirichlet boundary conditions, the analysis in [18] suggests that the eigenvectors do not form a Riesz basis. We have tried without success to modify the hypotheses of Theorem 5.1 to cover this case.

The conditions (5.10) and (5.11) say that the operator \mathcal{A}_2 in the diffusion equation is bounded in a certain sense with respect to the stiffness operator A_0 . We believe that some such relative boundedness is necessary for uniform exponential stability. A numerical experiment in which we used the one-dimensional wave equation for (2.10) and a fourth-order one-dimensional partial differential operator for \mathcal{A}_2 in (2.11) yielded a sequence of complex eigenvalues that appeared to approach the imaginary axis asymptotically.

6. An example and numerical results.

6.1. Linear model of a thermoelastic rod. We consider the axial vibrations of a visco-thermoelastic rod that is clamped and insulated at both ends. The length of the rod is normalized to 1. Control actuation is produced by a single force directed parallel to the rod and distributed uniformly over the rod segment $\eta_1 \leq \eta \leq \eta_2$. A sensor measures axial displacement at $\eta = \eta_1$ (i.e., the left end of the rod segment over which the actuator force is distributed). Finally we assume that both the actuator input and sensor output are corrupted by zero-mean Gaussian white noise with unit intensities.

The dynamics of the plant are described by the equations of one-dimensional linear thermoelasticity (see, for example, [7], [10], [33]), which consist of coupled one-dimensional wave and heat equations. If the rod has Kelvin–Voigt viscoelastic damping in addition to thermoelastic damping, then the state equations, boundary conditions, and output equation are

$$(6.1) \quad \rho \frac{\partial^2 w}{\partial t^2}(t, \eta) - \alpha_D(\lambda + 2\mu) \frac{\partial^3 w}{\partial \eta^2 \partial t}(t, \eta) - (\lambda + 2\mu) \frac{\partial^2 w}{\partial \eta^2}(t, \eta) \\ + \alpha_L(3\lambda + 2\mu) \frac{\partial \theta}{\partial \eta}(t, \eta) = b_0(\eta)u(t) + b_0(\eta)\gamma(t), \quad 0 < \eta < 1, \quad t > 0,$$

$$(6.2) \quad \rho c \frac{\partial \theta}{\partial t}(t, \eta) - \kappa \frac{\partial^2 \theta}{\partial \eta^2}(t, \eta) + \\ \bar{\theta} \alpha_L(3\lambda + 2\mu) \frac{\partial^2 w}{\partial \eta \partial t}(t, \eta) = 0, \quad 0 < \eta < 1, \quad t > 0,$$

$$(6.3) \quad w(t, 0) = 0 = w(t, 1), \quad t > 0,$$

$$(6.4) \quad \frac{\partial \theta}{\partial \eta}(t, 0) = 0 = \frac{\partial \theta}{\partial \eta}(t, 1), \quad t > 0,$$

$$(6.5) \quad y(t) = w(t, \eta_1) + \nu(t), \quad t > 0,$$

where w and θ are, respectively, the axial displacement and absolute temperature, ρ is the mass density, λ and μ are the Lamé (elasticity) parameters, c is the specific heat, and κ is the thermal conductivity. The positive constant $\bar{\theta}$ is a reference temperature—the absolute temperature of a stress-free reference state for the rod. The nonnegative

constants α_D and α_L are, respectively, the viscoelastic coefficient and the coefficient of thermal expansion, γ and ν are the noise processes, and the function $b_0 \in L_2(0, 1)$ is given by

$$(6.6) \quad b_0(\eta) = \begin{cases} 1, & \eta_1 \leq \eta \leq \eta_2 \\ 0, & \text{otherwise.} \end{cases}$$

Because of the insulated, or Neumann, boundary conditions in (6.4) on the temperature distribution, the open-loop system corresponding to (6.1)–(6.2) has a zero eigenvalue for which the associated eigenvector consists of zero displacement and velocity and nonzero uniform temperature distribution. This eigenvector is orthogonal (in $L_2(0, 1)$) to the control input function b_0 and is in the null space of the output operator corresponding to the measurement in (6.5), so that the span of this eigenvector is uncontrollable and unobservable. It follows that (i) the only part of the temperature distribution that can be controlled or observed is the part that is orthogonal to uniform temperature distributions; (ii) the average (over η) temperature in the rod, which we denote by θ_{ave} , is neither stabilizable nor detectable; (iii) θ_{ave} is a constant function of t .

Consequently, in the thermoelastic control problem, we replace the temperature distribution $\theta(t, \eta)$ with

$$(6.7) \quad \tilde{\theta}(t, \eta) = \theta(t, \eta) - \theta_{ave}.$$

The state equations, then, are (6.1)–(6.5) with θ replaced by $\tilde{\theta}$. The state space H has the structure in (2.20) with

$$(6.8) \quad H_0 = L_2(0, 1), \quad V_0 = H_0^1(0, 1),$$

$$(6.9) \quad H_2 = \{\phi \in L_2(0, 1) : \int_0^1 \phi d\eta = 0\}, \quad V_2 = H^1(0, 1) \cap H_2.$$

All of the spaces in this example are real. We use the standard L_2 inner product for H_0 , but we use

$$(6.10) \quad \langle \phi, \psi \rangle_2 = \frac{c}{\theta} \int_0^1 \phi \psi d\eta$$

for the inner product on H_2 . This inner product on H_2 is required to get the \mathcal{L}_0^* for which the semigroup generator in this example has the form in (2.19). For V_0 and V_2 , we use the norms

$$(6.11) \quad \|\phi\|_0 = \left(\int_0^1 |\phi'|^2 d\eta \right)^{1/2}, \quad \|\phi\|_2 = \left(\int_0^1 |\phi'|^2 d\eta \right)^{1/2}.$$

We define the operators $\mathcal{A}_j \in \mathcal{B}(V_j, V_j')$, $j = 0, 2$, $\mathcal{D}_0 \in \mathcal{B}(V_0, V_0')$, and $\mathcal{L}_0 \in \mathcal{B}(V_0, V_2')$ by

$$(6.12) \quad \langle \phi, \mathcal{A}_0 \psi \rangle_0 = \int_0^1 \frac{\lambda + 2\mu}{\rho} \phi' \psi' d\eta, \quad \phi, \psi \in V_0,$$

$$(6.13) \quad \langle \phi, \mathcal{A}_2 \psi \rangle_2 = \int_0^1 \frac{\kappa}{\rho \theta} \phi' \psi' d\eta, \quad \phi, \psi \in V_2,$$

$$(6.14) \quad \langle \phi, \mathcal{D}_0 \psi \rangle_0 = \int_0^1 \frac{\alpha_D(\lambda + 2\mu)}{\rho} \phi' \psi' d\eta, \quad \phi, \psi \in V_0,$$

and

$$(6.15) \quad \langle \phi, \mathcal{L}_0 \psi \rangle_2 = - \int_0^1 \frac{\alpha_L(3\lambda + 2\mu)}{\rho} \phi \psi' d\eta, \quad \phi \in V_2, \quad \psi \in V_0.$$

With these operators, the system in (6.1) and (6.2), with θ replaced by $\tilde{\theta}$, has the form in (2.22) with a semigroup generator of the form in (2.19).

From (6.12)–(6.15), it follows that we have all of the conditions in §5, including the hypotheses of Theorem 5.1 (assuming $\alpha_L > 0$; otherwise we would not have a thermoelastic problem). Hence, the open-loop thermoelastic system is uniformly exponentially stable, even if $\alpha_D = 0$.

For the numerical studies in this paper, we chose the parameters in (6.1) and (6.2) for an aluminum rod of length 100 in (see [27], [26]). With the length normalized to 1, the parameters take the values in Table 6.1.

TABLE 6.1
Parameters for (6.1)–(6.6).

$\rho = 9.82 \times 10^{-2}$	$\lambda = 2.064 \times 10^{-1}$	$\mu = 1.11 \times 10^{-1}$
$c = 5.40 \times 10^{-1}$	$\kappa = 7.02 \times 10^{-7}$	$\tilde{\theta} = 68$
$\alpha_L = 1.29 \times 10^{-3}$	$\alpha_D = 0$	
$\eta_1 = .385$	$\eta_2 = .486$	

The numerical results in this paper focus on the effects of thermoelastic damping. In [16] we presented numerical results for a similar example that included nonzero viscoelastic damping ($\alpha_D > 0$). The functional gains were much smoother than the gains for the case with thermoelastic damping only, and the approximating functional gains converged much faster. The numerical results in [16] indicate that, if Voigt–Kelvin viscoelastic damping is present, its effect dominates the effect of thermoelastic damping, but it is not clear whether Voigt–Kelvin viscoelastic damping is present at significant levels in common metals.

6.2. The optimal control problem and the approximation scheme. We have $m = \ell = p = 1$ with the input operators given by

$$(6.16) \quad B_0 r = \tilde{B}_0 r = \left(\frac{1}{\rho} b_0 \right) r, \quad r \in R^1, \quad B_2 = \tilde{B}_2 = 0,$$

and the output operator given by

$$(6.17) \quad C(\phi, \psi, \theta) = \phi(\eta_1), \quad (\phi, \psi, \theta) \in H = V_0 \times H_0 \times H_2.$$

In the quadratic performance index, we take the operator $Q \in \mathcal{B}(H)$ to be given by

$$(6.18) \quad Qx = Q(w, \dot{w}, \tilde{\theta}) = (w, \dot{w}, 0),$$

and we take $R = 1$. This Q penalizes the total mechanical energy in the rod but does not penalize temperature variations from the constant average value. The operator

$\hat{Q} \in \mathcal{B}(H)$ is given by (3.10) with $\tilde{B} = B$ given by (3.4) and (6.16). Since $\gamma(t)$ and $\nu(t)$ have unit intensities, $\Gamma = \hat{R} = 1$.

The functional control and estimator gains have the form $k_1 = (k_{1,1}, k_{1,2}, k_{1,3})$, and $\hat{k}_1 = (\hat{k}_{1,1}, \hat{k}_{1,2}, \hat{k}_{1,3})$ with $k_{1,1}, \hat{k}_{1,1} \in H_0^1(0, 1)$, $k_{1,2}, \hat{k}_{1,2} \in L_2(0, 1)$, and $k_{1,3}, \hat{k}_{1,3} \in H_2 \subset L_2(0, 1)$. If K and \hat{K} are, respectively, the control and estimator gain operators, then

$$(6.19) \quad Kx = \int_0^1 k'_{1,1} \phi' d\eta + \int_0^1 k_{1,2} \psi d\eta + \frac{c}{\theta} \int_0^1 k_{1,3} \theta d\eta, \quad x = (\phi, \psi, \theta) \in H,$$

and

$$(6.20) \quad \hat{K}r = (r\hat{k}_{1,1}, r\hat{k}_{1,2}, r\hat{k}_{1,3}) \in H, \quad r \in R^1.$$

In [16], we compared two Galerkin approximations for solving a linear-quadratic regulator problem for the thermoelastic rod in this example. One scheme was a finite element approximation in which linear splines were the basis vectors; in the other approximation, the open-loop eigenvectors of the distributed systems were the basis vectors. The modal approximation gave faster convergence for the approximating functional control gains. In this paper, we use the modal approximation only.

It is easy to see that, for the boundary conditions in this example, the eigenspaces of the open-loop thermoelastic rod are three-dimensional subspaces each spanned by a two-dimensional subspace of the undamped wave equation and a one-dimensional eigenspace of the heat equation. The eigenvectors of the wave equation are sine waves, and the eigenvectors of the heat equation are cosine waves. The sequence of three-dimensional subspaces of the thermoelastic rod are mutually orthogonal and complete in the state space H . Thus it is easy to show that all the conditions of Hypothesis 4.4 hold.

The open-loop eigenvalues can be determined as the solutions to the cubic characteristic equations corresponding to the three-dimensional eigenspaces. For the values of the parameters that we used, the eigenvalues corresponding to each open-loop subspace consist of a complex conjugate pair and a real eigenvalue, all with negative real parts. It can be shown by analysis of the sequence of cubic equations that, asymptotically, the real eigenvalues approach $-\infty$ and the complex pairs of eigenvalues approach a vertical line strictly to the left of the imaginary axis. This distribution of eigenvalues is not sufficient to guarantee (4.24); i.e., that the approximating open-loop semigroups are uniformly exponentially stable, with a decay rate uniform in n (the order of approximation, or number of modal subspaces). However, (4.24) does follow from the fact that the approximating open-loop semigroups used here are the projections onto modal subspaces of the original open-loop semigroup, which is uniformly exponentially stable according to Theorem 5.1. Hence (4.18) and (4.19) hold. Therefore, (4.20)–(4.30) are guaranteed.

To obtain the approximating control and estimator gains shown in Figs. 6.1 and 6.2, we used the matrix sign function method in [28] to solve Riccati matrix equations equivalent to the finite-dimensional Riccati operator equations (4.1) and (4.2), as discussed in §4.3. We used (4.74) and (4.75) with $m = p = 1$ to compute the approximating functional control gains $k_{1,i}^n$ ($i = 1, 2, 3$) and approximating functional estimator gains $\hat{k}_{1,i}^n$ ($i = 1, 2, 3$).

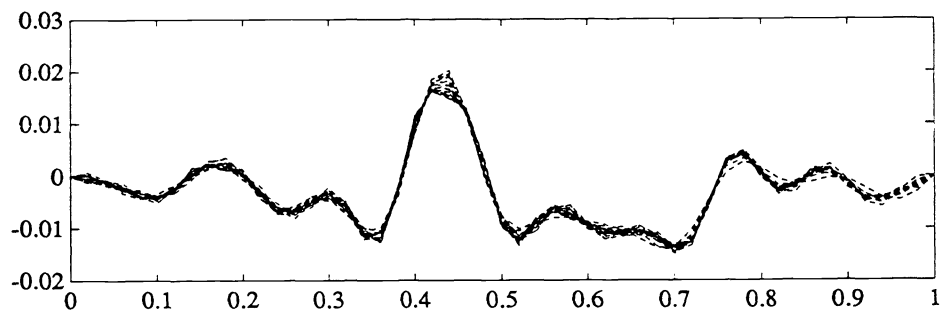


FIG. 6.1(a). Approximating functional control gains $k_{1,1}^n$, $n = 18, 19, \dots, 33$.

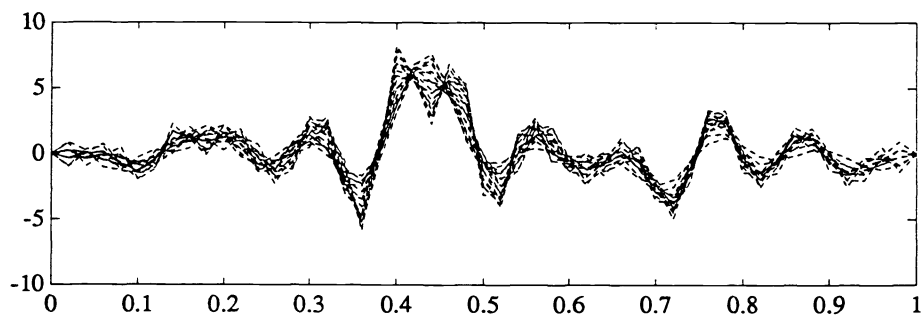


FIG. 6.1(b). Approximating functional control gains $k_{1,2}^n$, $n = 18, 19, \dots, 33$.

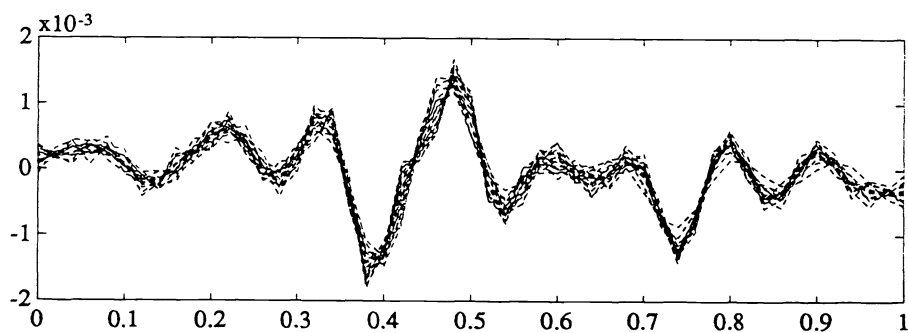


FIG. 6.1(c). Approximating functional control gains $k_{1,3}^n$, $n = 18, 19, \dots, 33$.

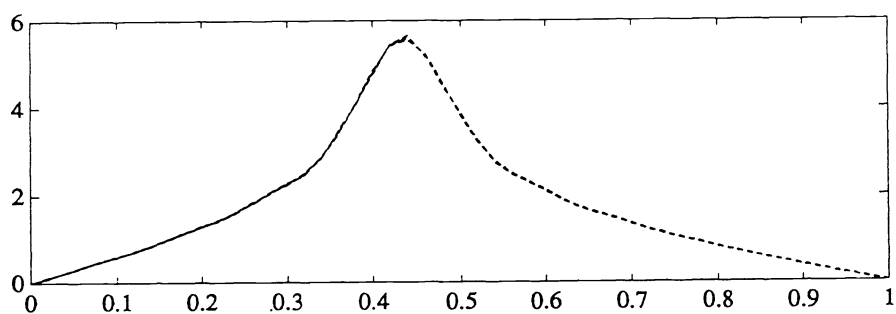


FIG. 6.2(a). Approximating functional estimator gains $\hat{k}_{1,1}^n$, $n = 18, 19, \dots, 33$.

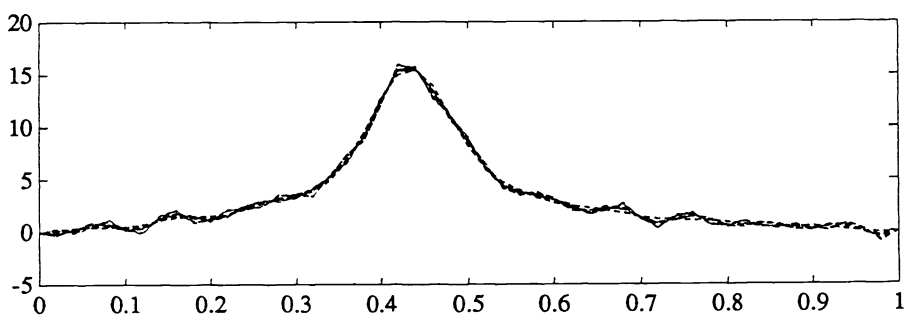


FIG. 6.2(b). Approximating functional estimator gains $\hat{k}_{1,2}^n$, $n = 18, 19, \dots, 33$.

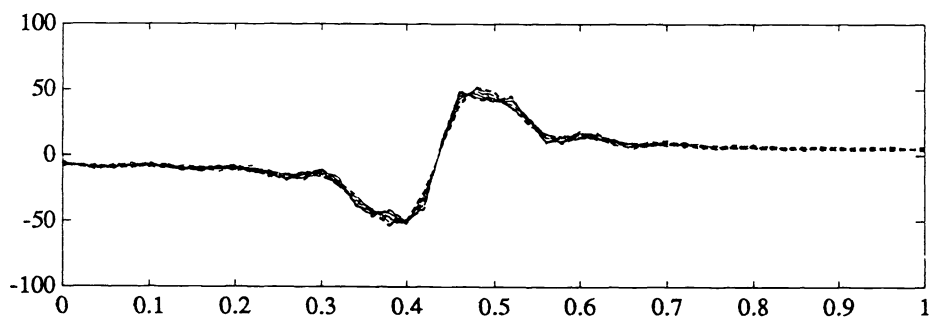


FIG. 6.2(c). Approximating functional estimator gains $\hat{k}_{1,3}^n$, $n = 18, 19, \dots, 33$.

6.3. Numerical results for finite-dimensional compensators. In each of the figures, we have plotted the approximation to the particular functional gain for each n between 18 and 33, where n is the number of modal subspaces used. Because the damping produced by thermoelastic dissipation is so small in this example, we see nothing resembling gain convergence until we use at least $n = 15$. The convergence results for approximations to the infinite-dimensional LQG problem guarantee that all of the functional gains do converge, but the convergence theory does not indicate the rate at which the gains converge. Numerical experience has shown that, generally, greater damping causes faster gain convergence.

We are not sure that we are seeing convergence in Fig. 6.1. Increasing n past 40 does not make the functional control gains look closer to any limit, and between $n = 40$ and $n = 50$, the numerical solution to the Riccati equation is so inaccurate in some cases that the corresponding gains do not resemble those in Fig. 6.1. While the functional gains must converge, it is possible that the order of approximation required for convergence exceeds our capability to solve the Riccati equations accurately. Another reason that we question whether our plots of the functional control gains show convergence is that when we compute the control gains for both $\alpha_D = 0$ and $\alpha_L = 0$, the plots look identical to Fig. 6.1. But with no damping for the wave equation and the coercive weighting that we place on the solution to the wave equation in the performance index, the norms of the finite-dimensional Riccati operators are guaranteed to grow without bound as n increases [12], [13]. Indeed, when $\alpha_D = 0$, and $\alpha_L = 0$, our numerical solutions to the Riccati equations break down for smaller n than they do when $\alpha_D = 0$ and $\alpha_L > 0$. There is some difference between the finite-dimensional gain matrices that we compute with and without thermoelastic damping in the plant model, but that difference is too small to be seen in plots of the functional gains.

The question arises, then, of whether the very light structural damping produced by the thermoelastic effect in the rod is significant in compensator design. To address this question, we computed eigenvalues for two closed-loop systems. Each closed-loop system was constructed by connecting a compensator based on a control model consisting of the first 20 modal subspaces to a simulation model, or truth model, consisting of the first 30 modal subspaces of the rod. Each compensator thus has dimension 60 while the simulation model has dimension 90. The 30-mode simulation model was the same in each case; it had the parameters in Table 6.1, including $\alpha_L = 1.29 \times 10^{-3}$. The 20-mode control model for Compensator 1 also had the parameters in Table 6.1. The control model for Compensator 2 had $\alpha_L = 0$, and all of the other parameters had the values in Table 6.1. This means that there is no damping for the mechanical vibrations of the rod in the open-loop control model for Compensator 2. Because the temperature distribution is not penalized in the performance index, the control gains $k_{1,3}$ and $k_{1,3}^n$ and estimator gains $\hat{k}_{1,3}$ and $\hat{k}_{1,3}^n$ are all zero in Compensator 2, and the gains $k_{1,1}$ and $k_{1,1}^n$, $\hat{k}_{1,1}$ and $\hat{k}_{1,1}^n$, $k_{1,2}$ and $k_{1,2}^n$, $\hat{k}_{1,2}$ and $\hat{k}_{1,2}^n$ are those that would be computed for a 20-mode model of the undamped wave equation alone.

Table 6.2 shows typical eigenvalues for the open-loop system and for the closed-loop system produced by each compensator. Since each compensator contains a copy of each of the first 20 modal subspaces, each closed-loop system contains six states, and six eigenvalues, corresponding to each of the first 20 modal subspaces. Each closed-loop system also contains the 30 states in twenty-first through thirtieth modal subspaces. While the closed-loop performance in the first ten or so modes is similar with both compensators, the closed-loop eigenvalues corresponding to several of the

TABLE 6.2
Typical open-loop and closed-loop eigenvalues.

Mode No.	Open-loop	Closed-loop with Compensator 1	Closed-loop with Compensator 2
1	$-2.30 \times 10^{-7} \pm i 6.57 \times 10^0$ -1.30×10^{-4}	$-3.15 \times 10^{-1} \pm i 6.57 \times 10^0$ $-1.44 \times 10^0 \pm i 6.89 \times 10^0$ -1.30×10^{-4} -1.30×10^{-4}	$-3.13 \times 10^{-1} \pm i 6.58 \times 10^0$ $-1.45 \times 10^0 \pm i 6.87 \times 10^0$ -1.30×10^{-4} -1.31×10^{-4}
2	$-9.19 \times 10^{-7} \pm i 1.31 \times 10^1$ -5.21×10^{-4}	$-1.26 \times 10^{-1} \pm i 1.31 \times 10^1$ $-1.22 \times 10^{-1} \pm i 1.31 \times 10^1$ -5.21×10^{-4} -5.21×10^{-4}	$-1.63 \times 10^{-1} \pm i 1.31 \times 10^1$ $-8.50 \times 10^{-2} \pm i 1.32 \times 10^1$ -5.21×10^{-4} -5.23×10^{-4}
3	$-2.07 \times 10^{-6} \pm i 1.97 \times 10^1$ -1.17×10^{-3}	$-2.55 \times 10^{-1} \pm i 1.97 \times 10^1$ $-3.45 \times 10^{-1} \pm i 1.97 \times 10^1$ -1.17×10^{-3} -1.17×10^{-3}	$-2.17 \times 10^{-1} \pm i 1.97 \times 10^1$ $-3.84 \times 10^{-1} \pm i 1.96 \times 10^1$ -1.18×10^{-3} -1.17×10^{-3}
10	$-2.30 \times 10^{-5} \pm i 6.57 \times 10^1$ -1.30×10^{-2}	$-1.82 \times 10^{-1} \pm i 6.57 \times 10^1$ $-8.03 \times 10^{-2} \pm i 6.57 \times 10^1$ -1.30×10^{-2} -1.30×10^{-2}	$-2.16 \times 10^{-1} \pm i 6.56 \times 10^1$ $-4.66 \times 10^{-2} \pm i 6.58 \times 10^1$ -1.30×10^{-2} -1.31×10^{-2}
14	$-4.50 \times 10^{-5} \pm i 9.20 \times 10^1$ -2.55×10^{-2}	$-3.44 \times 10^{-2} \pm i 9.20 \times 10^1$ $-3.41 \times 10^{-3} \pm i 9.20 \times 10^1$ -2.55×10^{-2} -2.55×10^{-2}	$-3.76 \times 10^{-2} \pm i 9.19 \times 10^1$ $-2.66 \times 10^{-4} \pm i 9.20 \times 10^1$ -2.56×10^{-2} -2.55×10^{-2}
18	$-7.45 \times 10^{-5} \pm i 1.18 \times 10^2$ -4.22×10^{-2}	$-1.53 \times 10^{-2} \pm i 1.18 \times 10^2$ $-2.16 \times 10^{-3} \pm i 1.18 \times 10^2$ -4.22×10^{-2} -4.22×10^{-2}	$-1.74 \times 10^{-2} \pm i 1.18 \times 10^2$ $-9.75 \times 10^{-5} \pm i 1.18 \times 10^2$ -4.22×10^{-2} -4.24×10^{-2}
19	$-8.30 \times 10^{-5} \pm i 1.25 \times 10^2$ -4.70×10^{-2}	$-1.03 \times 10^{-2} \pm i 1.25 \times 10^2$ $-1.98 \times 10^{-3} \pm i 1.25 \times 10^2$ -4.70×10^{-2} -4.70×10^{-2}	$-1.22 \times 10^{-2} \pm i 1.25 \times 10^2$ $-9.05 \times 10^{-5} \pm i 1.25 \times 10^2$ -4.70×10^{-2} -4.72×10^{-2}
20	$-9.19 \times 10^{-5} \pm i 1.31 \times 10^2$ -5.21×10^{-2}	$-2.51 \times 10^{-3} \pm i 1.31 \times 10^2$ $-6.67 \times 10^{-4} \pm i 1.31 \times 10^2$ -5.21×10^{-2} -5.21×10^{-2}	$-3.18 \times 10^{-3} \pm i 1.31 \times 10^2$ $-9.33 \times 10^{-5} \pm i 1.31 \times 10^2$ -5.21×10^{-2} -5.23×10^{-2}
21	$-1.01 \times 10^{-4} \pm i 1.38 \times 10^2$ -5.75×10^{-2}	$-1.04 \times 10^{-4} \pm i 1.38 \times 10^2$ -5.75×10^{-2}	$-1.04 \times 10^{-4} \pm i 1.38 \times 10^2$ -5.75×10^{-2}
22	$-1.11 \times 10^{-4} \pm i 1.45 \times 10^2$ -6.31×10^{-2}	$-1.34 \times 10^{-4} \pm i 1.45 \times 10^2$ -6.31×10^{-2}	$-1.34 \times 10^{-4} \pm i 1.45 \times 10^2$ -6.31×10^{-2}
Simulation Model: 30 Modal Subspaces, $\alpha_L = 1.29 \times 10^{-3}$ Compensator 1: 20 Modal Subspaces, $\alpha_L = 1.29 \times 10^{-3}$ Compensator 2: 20 Modal Subspaces, $\alpha_L = 0$			

higher-frequency modes reveal important differences between the two compensators. In particular, we note the second complex pair of closed-loop eigenvalues listed for mode 18. The magnitude of the real part produced by Compensator 1 is more than 20 times the corresponding number produced by Compensator 2. The same is true for mode 19. In certain high-frequency closed-loop states, then, the decay rates produced by Compensator 1 are more than 20 times the decay rates produced by Compensator 2.

The eigenvalues in Table 6.2 for modes 21 and 22, the first modes not modelled in the compensators, are typical of the eigenvalues for all ten modes that are present in the simulation model but not in the control models. These eigenvalues show that we have modelled enough modes in the compensators to eliminate any significant spillover between modelled and unmodelled modes.

Because the magnitudes of the real eigenvalues, which correspond to the heat equation (6.2), are so much larger than the magnitudes of the complex eigenvalues, we suspected that it might be possible to eliminate the states corresponding to the real open-loop eigenvalues from the control model and base a compensator design on a control model consisting of a sequence of second-order modes with eigenvalues equal to the complex open-loop eigenvalues of the thermoelastic rod. This amounts to putting artificial viscous damping in the wave equation.

We carried out such a design with twenty second-order modes having eigenvalues equal to the first twenty pairs of complex open-loop thermoelastic eigenvalues and mode shapes the same as the first twenty modes of the undamped rod. This compensator had dimension 40. When we closed the loop with the 30-mode simulation model used for Table 6.2 and computed the closed-loop eigenvalues, we obtained virtually identical results to those for Compensator 1, except that this third closed-loop system had only half as many real eigenvalues because the corresponding states were not modelled in the compensator. Even for modes 18 and 19, all of the closed-loop eigenvalues produced by the third compensator matched to at least three digits the corresponding eigenvalues produced by Compensator 1.

7. Conclusions. The abstract formulation of distributed models and the approximation theory developed in this paper apply to a wide variety of thermoelastic control systems. The uniform exponential stability result in §5 applies to a large class of thermoelastic problems, but not to certain systems that are known to be uniformly exponentially stable [20].

The numerical study in §6 focused on the effect of thermoelastic damping in optimal control of a flexible structure. The eigenvalue results demonstrate that, even though thermoelastic damping is small in common metals, a compensator based on a thermoelastic model of a flexible structure can produce significantly better response in high-frequency modes than a compensator based on an undamped model.

The theory in §2–4 also applies to thermoelastic control problems in which a thermal disturbance excites mechanical vibrations. This class of problems, which includes vibrations in flexible space structures caused by solar heating, might provide the most important applications for the theory developed here.

REFERENCES

- [1] A. BALAKRISHNAN, *Applied Functional Analysis*, 2nd ed., Springer-Verlag, New York, 1981.
- [2] H. BANKS AND K. ITO, *A unified framework for approximation and inverse problems for distributed parameter systems*, Control Theory and Advanced Technology, 4 (1988), pp. 73–90.
- [3] H. BANKS AND K. KUNISCH, *The linear regulator problem for parabolic systems*, SIAM J. Control Optim., 22 (1984), pp. 684–696.
- [4] J. BURNS, E. CLIFF, Z. LIU, AND R. MILLER, *Control of a thermoviscoelastic system*, in Proceedings of 27th IEEE Conference on Decision and Control, Austin, TX, December 1988, pp. 1249–1252.
- [5] J. BURNS, Z. LIU, AND R. MILLER, *Approximations of thermoelastic and viscoelastic control systems*, J. of Numer. Funct. Anal. Optim., 12 (1991), pp. 79–135.

- [6] J. BURNS, Z. LIU, AND S. ZHENG, *On the energy decay of a linear thermoelastic bar*, to appear.
- [7] D. CARLSON, *Linear thermoelasticity*, in Handbuch der Physik. Bd. VIa/2, C. Truesdell, ed., Springer-Verlag, Berlin, 1972.
- [8] R. CURTAIN AND A. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1978.
- [9] C. DAFERMOS, *On the existence and the asymptotic stability of solutions of the equations of linear thermoelasticity*, Arch. Rat. Mech. and Anal., 29 (1968), pp. 241–271.
- [10] A. DAY, *Heat Conduction within Linear Thermoelasticity*, Springer-Verlag, New York, 1985.
- [11] J. GIBSON, *The Riccati integral equations for optimal control problems on hilbert spaces*, SIAM J. Control Optim., 17 (1979), pp. 537–565.
- [12] ———, *An analysis of optimal modal regulation: Convergence and stability*, SIAM J. Control Optim., 19 (1981), pp. 686–707.
- [13] J. GIBSON AND A. ADAMIAN, *Approximation theory for LQG optimal control of flexible structures*, SIAM J. Control Optim., 29 (1991), pp. 1–37.
- [14] J. GIBSON AND I. ROSEN, *Computational methods for optimal linear-quadratic compensators for infinite dimensional discrete-time systems*, Lecture Notes in Control Inform. Sci., Vol. 135, Springer-Verlag, New York, Berlin, 1986, pp. 120–135.
- [15] ———, *Numerical approximation for the infinite-dimensional discrete-time optimal linear-quadratic regulator problem*, SIAM J. Control Optim., 26 (1988), pp. 428–451.
- [16] J. GIBSON, I. ROSEN, AND G. TAO, *Approximation in control of thermoelastic systems*, in Proc. 1989 American Control Conference, Pittsburgh, PA, June 1989, pp. 1171–1176.
- [17] ———, *Approximation in LQG optimal control of a thermoelastic rod*, in Proc. 3rd Annual Conference on Aerospace Computational Control, Oxnard, CA, August 1989.
- [18] S. HANSEN, *Exponential energy decay in a linear thermoelastic rod*, J. Math. Anal. Appl., to appear.
- [19] T. KATO, *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, New York, 1984.
- [20] J. KIM, *On the energy decay of a linear thermoelastic bar and plate*. Preprint, 1990.
- [21] J. LAGNESE, *Boundary stabilization of thin plates*, Studies in Applied Mathematics, Vol. 10, Society for Industrial and Applied Mathematics, Philadelphia, PA (1989).
- [22] ———, *The reachability problem for thermoelastic plates*, Arch. Rat. Mech. Anal., to appear.
- [23] Z. LIU, *Approximation and Control of a Thermoviscoelastic System*, Ph.D. thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA, 1989.
- [24] Z. LIU AND S. ZHENG, *Exponential stability of semigroup associated with thermoelastic system*, Quart. Appl. Math., to appear.
- [25] ———, *Uniform exponential stability and approximation in control of thermoelastic system*, to appear.
- [26] M. N. OZISIK, *Heat Conduction*, John Wiley, New York, 1980.
- [27] E. P. POPOV, *Introduction to Mechanics of Solids*, Prentice-Hall, Englewood Cliffs, NJ, 1968.
- [28] J. ROBERTS, *Linear model reduction and solution of the algebraic Riccati equation by use of the sign function*, Internat. J. of Control, 32 (1980), pp. 677–687.
- [29] I. ROSEN AND C. SU, *An approximation theory for the identification of linear thermoelastic systems*, J. Differential and Integral Equations, 4 (1991), pp. 783–802.
- [30] R. SHOWALTER, *Hilbert Space Methods for Partial Differential Equations*, Pitman, London, 1977.
- [31] M. SLEMROD, *Global existence, uniqueness, and asymptotic stability of classical smooth solutions in one-dimensional non-linear thermoelasticity*, Arch. Rat. Mech. Anal., 76 (1981), pp. 97–133.
- [32] H. TANABE, *Equations of Evolution*, Pitman, London, 1979.
- [33] J. WALKER, *Dynamical Systems and Evolution Equations*, Plenum, New York, 1980.
- [34] J. WALSH AND P. ULRICH, *Thermal blooming in the atmosphere*, in Laser Beam Propagation in the Atmosphere, J. Strohnbehn, ed., Springer-Verlag, Berlin, 1978, pp. 223–320.

ON MULTIVARIATE ROBUST STABILITY*

CHARLES K. CHUI[†] AND H. N. MHASKAR[‡]

Abstract. The notion of robust stability in the multidimensional setting is carefully investigated. In particular, the perturbation function is defined to measure the maximal stability region. Applications to robust stability tests of Kharitonov and Barmish types are obtained.

Key words. robust stability, perturbation functions, Kharitonov test

AMS(MOS) subject classifications. 93D05, 93C35, 93D20, 93D25

1. Introduction. In many applications to signal processing, systems theory, and optimal control, the problem of robust stability always arises. This paper is devoted to the study of maximal robust stability regions and robust stability tests for the multidimensional setting. The importance of multidimensional problems is well known, as documented by Bose, Zeheb, and others (cf. [2], [3], [5]). We give a careful formulation of multivariate robust stability and introduce the notion of the perturbation function generalizing the ideas in our earlier work [4] concerning the one-variable case. This function is then shown to determine the radius of a maximal stability ball in \mathbb{C}^N . The shape of the ball is determined by the norm under consideration. In particular, the coefficients need not vary linearly with the parameters. Our results are finer in the case of polydomains.

Applications to multivariate robust stability are also discussed in this paper. For a pathwise connected compact set, it suffices to check stability at one single point, judiciously chosen by means of the perturbation function. In addition, a Kharitonov-type test using the Minkowski functional is obtained. The stability test of Barmish [1] is generalized to the multivariate setting and to the case when the dependence of the coefficients on the parameters is nonlinear.

2. Notation. Let s be a positive integer and \mathbb{C}^s the s -dimensional Euclidean space consisting of vectors $\mathbf{z} = (z_1, \dots, z_s)$, where each z_k , $k = 1, \dots, s$, is a complex number. For $\mathbf{z} \in \mathbb{C}^s$, we use the notation

$$(2.1) \quad |\mathbf{z}| := \left(\sum_{k=1}^s |z_k|^2 \right)^{1/2}$$

for the Euclidean norm. For $\mathbf{z} \neq \mathbf{0}$, we define the projection of \mathbf{z} onto the unit sphere $\partial B_s := \{\mathbf{z}: |\mathbf{z}| = 1\}$ by

$$(2.2a) \quad \pi(\mathbf{z}) := \frac{\mathbf{z}}{|\mathbf{z}|}.$$

*Received by the editors July 30, 1990; accepted for publication (in revised form) June 21, 1991.

[†]Department of Electrical Engineering and Department of Mathematics, Texas A&M University, College Station, Texas 77843. This author's research was supported by SDIO/IST managed by Army Research Office contract numbers DAAL 03-87-K-0025 and DAAL-03-90-G-0091.

[‡]Department of Mathematics and Computer Science, California State University, Los Angeles, California 90032.

In general, for any set $A \subseteq \mathbb{C}^s$, we set

$$(2.2b) \quad \pi(A) := \{\pi(\mathbf{z}): \mathbf{z} \in A \setminus \{\mathbf{0}\}\},$$

and denote, by $\text{cl}(A)$ and ∂A , the closure and boundary of A , respectively. Hence, for the unit ball $B_s := \{\mathbf{z}: |\mathbf{z}| < 1\}$, its boundary is the unit sphere ∂B_s defined earlier. In addition, for any one-dimensional set $D \subseteq \mathbb{C}$, we will consider

$$D^s := \{(z_1, \dots, z_s): z_k \in D, k = 1, \dots, s\}.$$

In particular, if $U := B_1$ and $T := \partial B_1$ are the open unit disc and unit circle in \mathbb{C} , then U^s and T^s are the open unit polydisc and its distinguished boundary, respectively. We will also use the following standard multivariate notation: For $\mathbf{k} = (k_1, \dots, k_s)$, where each k_j , $j = 1, \dots, s$, is a nonnegative integer, and $\mathbf{z} = (z_1, \dots, z_s) \in \mathbb{C}^s$,

$$(2.3) \quad |\mathbf{k}| := k_1 + \dots + k_s, \quad \mathbf{z}^{\mathbf{k}} := z_1^{k_1} \dots z_s^{k_s}.$$

Our primary interest in this paper is the study of robust stability of polynomials in s complex variables. For this reason, any polynomial

$$\sum_{|\mathbf{k}| \leq n} a_{\mathbf{k}} \mathbf{z}^{\mathbf{k}},$$

of total degree n , is identified by its coefficients $\{a_{\mathbf{k}}\}$. To be more specific, we consider the N -tuple

$$\mathbf{a} = (a_{n,0,\dots,0}, \dots, a_{0,\dots,0,n}, a_{n-1,0,\dots,0}, \dots, a_{0,\dots,0,n-1}, \dots, a_{0,\dots,0}),$$

of these coefficients, where

$$(2.4) \quad N := \binom{n+s}{s},$$

and denote the above polynomial by

$$(2.5a) \quad P(\mathbf{a}, \mathbf{z}) := \sum_{|\mathbf{k}| \leq n} a_{\mathbf{k}} \mathbf{z}^{\mathbf{k}}.$$

We will also be interested in the homogeneous polynomial

$$(2.5b) \quad P_h(\mathbf{a}, \mathbf{z}) := \sum_{|\mathbf{k}|=n} a_{\mathbf{k}} \mathbf{z}^{\mathbf{k}} = \sum_{i_1+\dots+i_s=n} a_{i_1,\dots,i_s} z_1^{i_1} \dots z_s^{i_s},$$

which constitutes the leading terms of $P(\mathbf{a}, \cdot)$. It will often be convenient to think of the polynomial $P(\mathbf{a}, \mathbf{z})$ as the (complex) inner product of the vectors \mathbf{a} and (\mathbf{z}) where

$$(\mathbf{z}) := (\bar{z}_1^n, \bar{z}_1^{n-1} \bar{z}_2, \dots, \bar{z}_1^{n-1} \bar{z}_s, \dots, \bar{z}_1 \bar{z}_2^{n-1}, \dots, \bar{z}_1 \bar{z}_s^{n-1}, \dots, \bar{z}_s^n, \dots, \bar{z}_1, \dots, \bar{z}_s, 1).$$

Similarly, for the homogeneous polynomial, we introduce the notation

$$(2.6b) \quad (\mathbf{z}^h) := (\bar{z}_1^n, \bar{z}_1^{n-1} \bar{z}_2, \dots, \bar{z}_1^{n-1} \bar{z}_s, \dots, \bar{z}_1 \bar{z}_2^{n-1}, \dots, \bar{z}_1 \bar{z}_s^{n-1}, \dots, \bar{z}_s^n)$$

for any $\mathbf{z} = (z_1, \dots, z_s) \in \mathbb{C}^s$. Note that both s and n are fixed throughout this paper.

3. Stability and perturbation functions. Let $P(\mathbf{a}, \cdot)$ be a polynomial in s complex variables and E be an open set in \mathbb{C}^s . Intuitively, we would like to call $P(\mathbf{a}, \cdot)$ an E -stable polynomial if $P(\mathbf{a}, \cdot)$ is zero-free in $\mathbb{C}^s \setminus E$. This definition is indeed used in [5]. Given a norm $\|\cdot\|$ on \mathbb{C}^N , where N is defined in (2.4), we would then like to find the largest number δ (depending upon \mathbf{a}, E, n, s and the norm) such that whenever $\|\mathbf{b} - \mathbf{a}\| < \delta$ then $P(\mathbf{b}, \cdot)$ is also E -stable. The following example shows that when $\mathbb{C}^s \setminus E$ is unbounded, the notion of stability needs further careful refinement for this effort to be successful.

Example 3.1. Let $s = 2, n = 1, E := \{(z_1, z_2) : \operatorname{Re} z_1 < 0\}$, $\mathbf{a} := (1, 0, 1)$ so that $P(\mathbf{a}, \mathbf{z}) = z_1 + 1$ is an E -stable polynomial. Every ball around \mathbf{a} contains vectors of the form $\mathbf{a}_m := (1, 1/m, 1)$ for all sufficiently large integers m . However, the polynomial $P(\mathbf{a}_m, \mathbf{z}) = z_1 + z_2/m + 1$ is clearly not E -stable for any positive integer m .

We observe that a critical aspect here is, loosely speaking, the wandering-off of certain zeros to infinity. The following definition is motivated by the desire to control this phenomenon.

DEFINITION 3.2. Let $s \geq 1$ and $n \geq 1$ be integers, and E be an open set in \mathbb{C}^s . A polynomial $P(\mathbf{a}, \cdot)$, or equivalently, its coefficient vector \mathbf{a} , will be called E -stable (or more precisely when needed, (E, n) -stable) if each of the following conditions is satisfied:

- (i) If $\mathbf{z} \in \mathbb{C}^s$ and $P(\mathbf{a}, \mathbf{z}) = 0$ then $\mathbf{z} \in E$.
- (ii) Suppose that $\mathbb{C}^s \setminus E$ is an unbounded set. Then, in addition to (i), $P_h(\mathbf{a}, \mathbf{z}) \neq 0$ for all $\mathbf{z} \in \operatorname{cl}(\pi(\mathbb{C}^s \setminus E))$.

For the one-variable setting when $P_h(z) = a_0 z^n$, condition (ii) for unbounded $\mathbb{C} \setminus E$ above is equivalent to the condition that the leading coefficient a_0 is nonzero (i.e., $P(\mathbf{a}, \cdot)$ is of precise degree n). Thus, Definition 3.2 agrees with Definition 2.1 in our earlier work [4] (cf. Theorem 3.3 therein).

We emphasize that the “adjustment” to the intuitive notion of stability is required only when $\mathbb{C}^s \setminus E$ is unbounded.

Next, we define the perturbation function. For $\mathbf{z} \in \mathbb{C}^s$, we let

$$(3.1a) \quad \Pi_{\mathbf{z}} := \{\mathbf{a} \in \mathbb{C}^N : P(\mathbf{a}, \mathbf{z}) = 0\};$$

$$(3.1b) \quad \Pi_{\mathbf{z}}^h := \{\mathbf{a} \in \mathbb{C}^N : P_h(\mathbf{a}, \mathbf{z}) = 0\}.$$

Furthermore, for any norm $\|\cdot\|$ defined on \mathbb{C}^N , we use the notation $d(\mathbf{a}, A)$ for the distance in this norm between any $\mathbf{a} \in \mathbb{C}^N$ to a closed set $A \subseteq \mathbb{C}^N$, namely,

$$(3.2) \quad d(\mathbf{a}, A) := \inf_{\mathbf{b} \in A} \|\mathbf{a} - \mathbf{b}\|.$$

If E is an open set in \mathbb{C}^s , then we also set

$$(3.3a) \quad d(\mathbf{a}) := d(\mathbf{a}; E, n, s, \|\cdot\|) := \inf_{\mathbf{z} \in \partial E} d(\mathbf{a}, \Pi_{\mathbf{z}});$$

$$(3.3b) \quad d_{\infty}(\mathbf{a}) := d_{\infty}(\mathbf{a}; E, n, s, \|\cdot\|) := \inf_{\mathbf{z} \in \operatorname{cl}(\pi(\mathbb{C}^s \setminus E))} d(\mathbf{a}, \Pi_{\mathbf{z}}^h).$$

DEFINITION 3.3. Let $\mathbf{a} \in \mathbb{C}^N$ and E be an open set in \mathbb{C}^s . The perturbation function $\delta: \mathbb{C}^N \rightarrow [0, \infty)$ is defined by

$$(3.4) \quad \begin{aligned} \delta(\mathbf{a}) &:= \delta(\mathbf{a}; E, n, s, \|\cdot\|) \\ &:= \begin{cases} d(\mathbf{a}; E, n, s, \|\cdot\|), & \text{if } \mathbb{C}^s \setminus E \text{ is compact,} \\ \min\{d(\mathbf{a}; E, n, s, \|\cdot\|), d_{\infty}(\mathbf{a}; E, n, s, \|\cdot\|)\}, & \text{otherwise.} \end{cases} \end{aligned}$$

The importance of the perturbation function will be clear from the following result.

THEOREM 3.4. *Let $E \subseteq \mathbb{C}^s$ be open, and $\mathbf{a}^* \in \mathbb{C}^N$ be E -stable. Also, let δ be the perturbation function defined as in (3.4). Then*

- (i) $\delta(\mathbf{a}^*) > 0$;
- (ii) $\|\mathbf{b} - \mathbf{a}^*\| < \delta(\mathbf{a}^*)$ implies \mathbf{b} is E -stable; and
- (iii) there exists $\hat{\mathbf{b}} \in \mathbb{C}^N$ which is not E -stable but $\|\hat{\mathbf{b}} - \mathbf{a}^*\| = \delta(\mathbf{a}^*)$.

We observe that Theorem 3.4 is a precise statement of the continuous dependence of the zeros of a polynomial on its coefficients.

The quantity $\delta(\mathbf{a}^*)$ is thus the multivariate analogue of the perturbation constant defined in our earlier work [4]. We observe that an application of the Hahn–Banach Theorem yields the following alternative expressions for the distances $d(\mathbf{a}, \Pi_{\mathbf{z}})$ and $d(\mathbf{a}, \Pi_{\mathbf{z}}^h)$, namely,

$$(3.5a) \quad d(\mathbf{a}, \Pi_{\mathbf{z}}) = \frac{|P(\mathbf{a}, \mathbf{z})|}{\|(\mathbf{z})\|^*};$$

$$(3.5b) \quad d_{\infty}(\mathbf{a}, \Pi_{\mathbf{z}}^h) = \frac{|P_h(\mathbf{a}, \mathbf{z})|}{\|(\mathbf{z}^h)\|^*},$$

where (\mathbf{z}) and (\mathbf{z}^h) are defined in (2.6) and $\|\cdot\|^*$ is the dual norm corresponding to $\|\cdot\|$, in the sense that

$$(3.6) \quad \|\mathbf{c}\|^* := \sup_{\|\mathbf{c}'\|=1} \left| \sum_{k=1}^s c_k \bar{c}'_k \right|, \quad \mathbf{c} \in \mathbb{C}^N.$$

Thus, we also have

$$(3.7) \quad \delta(\mathbf{a}; E, n, s, \|\cdot\|) = \begin{cases} \inf_{\mathbf{z} \in \partial E} \frac{|P(\mathbf{a}, \mathbf{z})|}{\|(\mathbf{z})\|^*}, & \text{if } \mathbb{C}^s \setminus E \text{ is compact,} \\ \min \left\{ \inf_{\mathbf{z} \in \partial E} \frac{|P(\mathbf{a}, \mathbf{z})|}{\|(\mathbf{z})\|^*}, \inf_{\mathbf{z} \in \text{cl}(\pi(\mathbb{C}^s \setminus E))} \frac{|P_h(\mathbf{a}, \mathbf{z})|}{\|(\mathbf{z}^h)\|^*} \right\}, & \text{otherwise.} \end{cases}$$

Of course, Theorem 3.4 is the multivariate extension of Theorem 3.1 of our earlier work [4]. However, we do not know at this stage if $d(\mathbf{a}) \leq d_{\infty}(\mathbf{a})$ in general, even when the norm $\|\cdot\|$ is quasi-monotone in the sense discussed in [4].

Under special circumstances, we may replace the expression

$$\inf_{\mathbf{z} \in \partial E} \frac{|P(\mathbf{a}, \mathbf{z})|}{\|(\mathbf{z})\|^*}$$

in (3.7) by another expression where the infimum is taken over a substantially smaller set than ∂E .

Let $\varphi: \text{cl}(U) \rightarrow \mathbb{C} \cup \{\infty\}$ be a function holomorphic in U and continuous on $\text{cl}(U)$. Also, let $F := \varphi(\text{cl } U)$ and

$$(3.8) \quad E := \mathbb{C}^s \setminus F^s.$$

In view of the open mapping theorem, we have $\partial F \subseteq \varphi(T)$, so that the distinguished boundary of F^s is contained in $\varphi(T)^s$. In many applications, φ would be one-one and $\varphi(T)^s$ would be the distinguished boundary of F^s .

In the following theorem, we need this notation:

$$(3.9) \quad d_2(\mathbf{a}^*) := \inf_{\mathbf{z} \in \varphi(T)^s} \frac{|P(\mathbf{a}^*, \mathbf{z})|}{\|\mathbf{z}\|^*}$$

and

$$(3.10) \quad \delta'(\mathbf{a}^*) := \begin{cases} d_2(\mathbf{a}^*), & \text{if } F^s \text{ is compact;} \\ \min(d_2(\mathbf{a}^*), d_\infty(\mathbf{a}^*)), & \text{otherwise.} \end{cases}$$

THEOREM 3.5. *Let E be an open set as defined in (3.8), and let \mathbf{a}^* be E -stable. Then*

- (i) $\delta'(\mathbf{a}^*) > 0$;
- (ii) $\|\mathbf{b} - \mathbf{a}^*\| < \delta'(\mathbf{a}^*)$ implies \mathbf{b} is E -stable; and
- (iii) there exists $\hat{\mathbf{b}} \in \mathbb{C}^N$ which is not E -stable but $\|\hat{\mathbf{b}} - \mathbf{a}^*\| = \delta'(\mathbf{a}^*)$.

4. Applications to robust stability tests. Let $K \subseteq \mathbb{C}^N$ and E be an open set in \mathbb{C}^s . We say that K is E -stable if every \mathbf{a} in K is E -stable. In this section, we apply the results in §3 to develop several necessary and sufficient conditions for the set K to be E -stable. The following result is fairly general.

THEOREM 4.1. *Let K be a pathwise connected set in \mathbb{C}^N and K contain at least one E -stable vector. Then K is E -stable if and only if the perturbation function $\delta(\mathbf{a}; E, n, s, \|\cdot\|)$ is positive for every $\mathbf{a} \in K$ and some norm $\|\cdot\|$ on \mathbb{C}^N . If K is compact, then there exists an $\mathbf{a}^* := \mathbf{a}^*(E, n, s, \|\cdot\|, K) \in K$ with the property that K is E -stable if and only if \mathbf{a}^* is E -stable.*

We emphasize that Theorem 4.1 is purely a qualitative result; the choice of the norm is irrelevant. The location of the point \mathbf{a}^* will generally depend on the norm.

The following theorem is similar in spirit to the celebrated Kharitonov theorem and extends Theorem 3.4 of [4]. We recall that if \mathcal{B} is a compact, convex, balanced, absorbing set in \mathbb{C}^N , then the Minkowski functional of \mathcal{B} , defined by

$$(4.1) \quad \|\mathbf{a}\|_{\mathcal{B}} := \inf\{t: t^{-1}\mathbf{a} \in \mathcal{B}\},$$

is a norm on \mathbb{C}^N and that $\mathcal{B} = \{\mathbf{a} \in \mathbb{C}^N: \|\mathbf{a}\|_{\mathcal{B}} \leq 1\}$, (cf. [8]).

THEOREM 4.2. *Let $\mathbf{a}^* \in \mathbb{C}^N$, \mathcal{B} be a compact, convex, balanced, and absorbing set in \mathbb{C}^N , $K := \mathcal{B} + \mathbf{a}^*$ and $\|\cdot\| = \|\cdot\|_{\mathcal{B}}$ be as in (4.1). Then K is E -stable if and only if \mathbf{a}^* is E -stable and $\delta(\mathbf{a}^*; E, n, s, \|\cdot\|) > 1$.*

It is important to note that the set \mathcal{B} need not be a polytope, so that when the coefficients vary depending upon certain parameters, this dependence can well be fairly complex. In particular, the coefficients need not vary independently of one another. Furthermore, depending on the set \mathcal{B} , the actual formula for the Minkowski functional may be fairly complicated. The theorem does lead to fairly simple criteria in some special cases. For instance, if $E = \mathbb{C}^s \setminus cl(U^s)$ and the set K is of the form $\|a - a^*\|_p \leq \delta$, then the stability test is simply

$$\min_{\mathbf{z} \in T^s} |P(a^*, z)| > \delta N^{(1-p)/p}.$$

The evaluation of the minimum expression can be done by a simple sweep over the distinguished boundary T^s .

In the following, we consider the coefficient space to be \mathbb{R}^{2N} rather than \mathbb{C}^N . If $K \subseteq \mathbb{R}^{2N}$ is compact and convex, then it is well known that there are at most $2N + 1$ points in K such that every element of K is a convex combination of these points.

These points will be called the vertices of the set K . Any set of vertices of K will be denoted by V_K . If K is a polytope, then the set of vertices of K in the usual sense may indeed serve as V_K .

Let $E \subseteq \mathbb{C}^s$ be open and K be a compact convex set in \mathbb{R}^{2N} . We present in the following theorem an analogue of a stability test due to Barmish [1]. Let $E \subseteq \mathbb{C}^s$ be open and Γ be any curve in \mathbb{R}^2 surrounding $\mathbf{0}$. Also, let

$$(4.2a) \quad b_1(K, E, \Gamma) := \inf_{\mathbf{a} \in V_K, \mathbf{z} \in \partial E} \sup_{(\lambda, \mu) \in \Gamma} \{ \lambda \operatorname{Re} P(\mathbf{a}, \mathbf{z}) + \mu \operatorname{Im} P(\mathbf{a}, \mathbf{z}) \}$$

and

$$(4.2b) \quad b_\infty(K, E, \Gamma) := \inf_{\mathbf{a} \in V_K, \mathbf{z} \in \operatorname{cl}(\pi(\mathbb{C}^s \setminus E))} \sup_{(\lambda, \mu) \in \Gamma} \{ \lambda \operatorname{Re} P(\mathbf{a}, \mathbf{z}) + \mu \operatorname{Im} P(\mathbf{a}, \mathbf{z}) \}.$$

We have the following result.

THEOREM 4.3. *Let $E, K, \Gamma, b_1, b_\infty$ be as defined above such that K contains at least one E -stable element.*

(i) *Suppose that $\mathbb{C}^s \setminus E$ is bounded. Then K is E -stable if and only if*

$$b_1(K, E, \Gamma) > 0.$$

(ii) *Suppose that $\mathbb{C}^s \setminus E$ is unbounded. Then K is E -stable if and only if both*

$$b_1(K, E, \Gamma) > 0$$

and

$$b_\infty(K, E, \Gamma) > 0.$$

(iii) *If $\mathbb{C}^s \setminus E$ is a polydomain satisfying the hypotheses in Theorem 3.5, then the infimum over ∂E in (4.2a) can be replaced by the infimum over $\varphi(T)^s$.*

We observe that the computations involved in (4.2a) and (4.2b) can be somewhat simplified by taking Γ to be the curve $|\lambda| + |\mu| = 1$. In any case, the infimum over ∂E and $\operatorname{cl}(\pi(\mathbb{C}^s \setminus E))$ are the only difficult extremal values to calculate. If K is not a polytope, then finding a proper vertex set might be a problem as well. The connection between Kharitonov's test and Barmish's is already explained in [1].

5. Examples. In this section, we give some examples to illustrate Theorems 3.4 and 3.5.

First, we observe that when $E = \mathbb{C}^s \setminus \operatorname{cl}(U^s)$ and the ℓ^p -norm is considered, Theorem 3.5 shows that the perturbation constant for an E -stable polynomial P is given by

$$(5.1) \quad N^{\frac{1}{p}-1} \min_{|z_1|=\dots=|z_2|=1} |P(\mathbf{z})|.$$

Example 5.1. Let $E = \mathbb{C}^2 \setminus \operatorname{cl}(U^2)$,

$$(5.2) \quad P(z_1, z_2) := (3 + 0.6z_2) + (3.5 + 0.7z_2)z_1 + (1 + 0.2z_2)z_1^2$$

and consider the coefficient variation in the ℓ^∞ -ball. The perturbation constant in this case is given by (5.1), namely,

$$(5.3) \quad \begin{aligned} \min_{|z_1|=|z_2|=1} \frac{|P(z_1, z_2)|}{10} \\ = \frac{1}{100} \min_{|z_1|=|z_2|=1} |(2 + z_1)(3 + 2z_1)(5 + z_2)| \\ = 0.04. \end{aligned}$$

Example 5.2 (cf. [5]). Let $E = \mathbb{C}^2 \setminus \{(z_1, z_2) : \operatorname{Re} z_1 \geq 0, \operatorname{Re} z_2 \geq 0\}$ and

$$(5.4) \quad P(z_1, z_2) = 8 + 4z_1 + 5z_2 + 3z_1z_2.$$

We observe that P is not E -stable in the sense of Definition 3.2. Nevertheless, the ideas in the proof of Theorem 3.4 help us calculate the maximal value of δ so that $a_0 + a_1z_2 + a_2z_2 + a_3z_1z_2$ is Hurwitz in the sense of [5] for all values of (a_0, a_1, a_2, a_3) satisfying

$$|a_0 - 8| + |a_1 - 4| + |a_2 - 5| + |a_3 - 3| \leq \delta.$$

This δ is given by

$$(5.5) \quad \delta = \min_{x, y \in \mathbb{R}} \frac{|8 - 3xy + i(4x + 5y)|}{\max\{1, |x|, |y|, |xy|\}}.$$

It is elementary to verify that when a, b, c, d are positive numbers, and $2bc > ad$, then

$$(5.6) \quad \min_{x, y \in \mathbb{R}} (a - dxy)^2 + (bx + cy)^2 = a^2.$$

Hence, it follows that

$$(5.7) \quad \begin{aligned} \delta^2 = \min\{ & \min_{|x|, |y| \leq 1} (8 - 3xy)^2 + (4x + 5y)^2, \\ & \min_{|x|, |y| \leq 1} (4 - 5xy)^2 + (8x + 3y)^2, \\ & \min_{|x|, |y| \leq 1} (5 - 4xy)^2 + (8x + 3y)^2, \\ & \min_{|x|, |y| \leq 1} (3 - 8xy)^2 + (4x + 5y)^2\} = 9. \end{aligned}$$

This coincides with the value calculated in [5] for the ℓ^∞ -ball, except that now the coefficients are considered complex and a linear dependence on parameters is considered. In any case, our method involves only a sweep over the boundary in contrast to the evaluation of certain determinants and solutions to inequalities as in [5].

Next, we illustrate an application of Theorem 3.4 when $E = \mathbb{C}^s \setminus B_s$. It is convenient to consider, instead of spherical regions, the following ellipsoidal stability balls

$$(5.8) \quad \sum_{|\mathbf{k}| \leq n} |a_{\mathbf{k}} - a_{\mathbf{k}}^*|^2 / \binom{|\mathbf{k}|}{\mathbf{k}!} \leq \delta.$$

The formula (3.7) for the perturbation constant in this case becomes

$$(5.9) \quad \delta(\mathbf{a}^*) = \min_{|\mathbf{z}|=1} |P(\mathbf{a}^*, \mathbf{z})| / \sqrt{n+1}.$$

This is used in the following example.

Example 5.3. Let $E = \mathbb{C}^2 \setminus B_2$ and

$$(5.10) \quad P(z_1, z_2) = 2 + z_1z_2.$$

Writing $\mathbf{a}^* = (0, 1, 0, 0, 0, 2)$, we seek the maximum stability ball of the form

$$(5.11) \quad (a_5 - 2)^2 + a_4^2 + a_3^2 + a_2^2 + \frac{1}{2}(a_1 - 1)^2 + a_0^2 \leq \delta^2.$$

The perturbation constant δ is then given by Theorem 3.5 as follows:

$$(5.12) \quad \delta = \min_{|z_1|^2 + |z_2|^2 = 1} \frac{|P(z_1, z_2)|}{(1 + |z_1|^2 + |z_2|^2 + |z_1|^4 + 2|z_1|^2|z_2|^2 + |z_2|^4)^{1/2}} \\ = \frac{1}{\sqrt{3}} \min_{|z_1|^2 + |z_2|^2 = 1} |P(z_1, z_2)|.$$

Since

$$\max_{|z_1|^2 + |z_2|^2 = 1} |z_1 z_2| = \frac{1}{2},$$

we see that

$$\delta = \frac{1}{\sqrt{3}} \left(2 - \frac{1}{2} \right) = \frac{\sqrt{3}}{2}.$$

6. Proof of the results. The following well-known result is central to the proofs of most of the theorems in this paper.

PROPOSITION 6.1 ([6, pg. 272]). *Let Ω be an open set in \mathbb{C}^s , $\{f_m\}_{m=1}^\infty$ a sequence of functions holomorphic and zero-free in Ω , f a nontrivial function holomorphic in Ω , and let $f_m \rightarrow f$ uniformly on every compact subset of Ω . Then f is also zero-free in Ω .*

Using Proposition 6.1, we first prove that when $E \subseteq \mathbb{C}^s$ is open, the set of all E -stable polynomials is open.

PROPOSITION 6.2. *Let E be an open set in \mathbb{C}^s and $n \geq 1$ be a fixed integer. Then the collection of all E -stable polynomials is an open set in the topology of uniform convergence on compact subsets of \mathbb{C}^s .*

Proof of Proposition 6.2. Let $\{P^{(m)}\}$ be a sequence of polynomials of degree $\leq n$ that are not E -stable, P be a polynomial of degree $\leq n$, and $P^{(m)} \rightarrow P$ uniformly on compact subsets of \mathbb{C}^s . If each $P^{(m)}$ has a zero $\mathbf{w}^{(m)} \in \mathbb{C}^s \setminus E$ and the sequence $\{\mathbf{w}^{(m)}\}$ has a limit point $\mathbf{w} \in \mathbb{C}^s \setminus E$, then it is easy to see that $P(\mathbf{w}) = 0$ and hence P is not E -stable. If the sequence $\{\mathbf{w}^{(m)}\}$ does not have a limit point, then necessarily $\mathbb{C}^s \setminus E$ is unbounded and $|\mathbf{w}^{(m)}| \rightarrow \infty$ as $m \rightarrow \infty$. We observe that the sequence $\{P^{(m)}(\mathbf{z})/(1 + |\mathbf{z}|)^n\}$ converges to $P(\mathbf{z})/(1 + |\mathbf{z}|)^n$ uniformly on the whole space \mathbb{C}^s . Hence, in view of the fact that $|\mathbf{w}^{(m)}| \rightarrow \infty$, and $P^{(m)}(\mathbf{w}^{(m)}) = 0$, $m = 1, 2, \dots$, we have

$$\begin{aligned} \lim_{m \rightarrow \infty} |P_h(\pi(\mathbf{w}^{(m)}))| &= \lim_{m \rightarrow \infty} |P_h(\mathbf{w}^{(m)})|/|\mathbf{w}^{(m)}|^n \\ &= \lim_{m \rightarrow \infty} |P(\mathbf{w}^{(m)})|/|\mathbf{w}^{(m)}|^n \\ &= \lim_{m \rightarrow \infty} \frac{|P(\mathbf{w}^{(m)})|}{(1 + |\mathbf{w}^{(m)}|)^n} = \lim_{m \rightarrow \infty} \frac{|P^{(m)}(\mathbf{w}^{(m)})|}{(1 + |\mathbf{w}^{(m)}|)^n} \\ &= 0. \end{aligned}$$

If \mathbf{w} is a limit point of $\pi(\mathbf{w}^{(m)})$, then $\mathbf{w} \in \text{cl}(\pi(\mathbb{C}^s \setminus E))$ and $P_h(\mathbf{w}) = 0$. Thus P is not E -stable. When $\mathbb{C}^s \setminus E$ is unbounded, we need to consider another possibility as follows. If $\{P^{(m)}\}$ has a subsequence $\{Q^{(k)}\}$ such that each $Q^{(k)}$ has a zero $\mathbf{z}^{(k)} \in \text{cl}(\pi(\mathbb{C}^s \setminus E))$, then we observe that $Q_h^{(k)} \rightarrow P_h$ uniformly on $|\mathbf{z}| = 1$, that $\{\mathbf{z}^{(k)}\}$ has a limit point $\mathbf{z} \in \text{cl}(\pi(\mathbb{C}^s \setminus E))$, and finally that $P_h(\mathbf{z}) = 0$. Thus, P is not E -stable. \square

Proof of Theorem 3.4. We present the proof in the case when $\mathbb{C}^s \setminus E$ is unbounded, since the other case is similar and simpler. Let $r := \min(d(\mathbf{a}^*), d_\infty(\mathbf{a}^*))$ and $\|\mathbf{a} - \mathbf{a}^*\| < r$, and assume that \mathbf{a} is not E -stable. Let

$$t := \sup\{\alpha \in [0, 1]: \alpha \mathbf{a} + (1 - \alpha) \mathbf{a}^* \text{ is } E\text{-stable}\},$$

and $\mathbf{c} := t\mathbf{a} + (1-t)\mathbf{a}^*$. Then $\|\mathbf{c} - \mathbf{a}^*\| < r$. In view of Proposition 6.2, \mathbf{c} is not E -stable either. Moreover, it follows from Proposition 6.1 applied with $\mathbb{C}^s \setminus \text{cl}(E)$ in place of Ω that $P(\mathbf{c}, \cdot)$, being the limit of polynomials zero-free in $\mathbb{C}^s \setminus \text{cl}(E)$, is either identically zero or zero-free in $\mathbb{C}^s \setminus \text{cl}(E)$. Hence, either $P(\mathbf{c}, \cdot)$ has a zero on ∂E or $P_h(\mathbf{c}, \cdot)$ has a zero on $\text{cl}(\pi(\mathbb{C}^s \setminus E))$. In either case, $\|\mathbf{a}^* - \mathbf{c}\| \geq r$, which is a contradiction.

Next, let $r = d(\mathbf{a}^*)$. We choose a sequence $\{\mathbf{w}^{(m)}\}$ in ∂E and $\mathbf{a}^{(m)} \in \Pi_{\mathbf{w}^{(m)}}$ such that $r \leq \|\mathbf{a} - \mathbf{a}^{(m)}\| \leq r + 1/m$. If $\mathbf{w}^{(m)}$ has a limit point $\mathbf{w} \in \partial E$ (which would necessarily be the case if $\mathbb{C}^s \setminus E$ is bounded), then $\{\mathbf{a}^{(m)}\}$ would have a limit point \mathbf{b} with the property that $P(\mathbf{b}, \mathbf{w}) = 0$. So, necessarily, $\|\mathbf{a}^* - \mathbf{b}\| = r$ and \mathbf{b} is not E -stable. If $\mathbf{w}^{(m)}$ does not have a limit point, then $|\mathbf{w}^{(m)}| \rightarrow \infty$. Let \mathbf{b} be a limit point of the sequence $\{\mathbf{a}^{(m)}\}$. Then, as in the proof of Proposition 6.2, $P_h(\mathbf{b}, \mathbf{z})$ would have a zero on $\text{cl}(\pi(\mathbb{C}^s \setminus E))$. So $P(\mathbf{b}, \cdot)$ is not E -stable and $\|\mathbf{a}^* - \mathbf{b}\| = r$.

If $r = d_\infty(\mathbf{a}^*)$, then we may select $\mathbf{z}^{(m)} \in \text{cl}(\pi(\mathbb{C}^s \setminus E))$ and $\mathbf{a}^{(m)} \in \Pi_{\mathbf{z}^{(m)}}^h$ such that $\|\mathbf{a}^* - \mathbf{a}^{(m)}\| \leq r + 1/m$. If $\mathbf{z}^{(0)}$ is a limit point of $\{\mathbf{z}^{(m)}\}$, then $\{\mathbf{a}^{(m)}\}$ has a limit point $\mathbf{b} \in \Pi_{\mathbf{z}^{(0)}}^h$. Thus, \mathbf{b} is not E -stable and $\|\mathbf{a}^* - \mathbf{b}\| = r$.

In view of Proposition 6.2, r is necessarily positive. \square

The proof of Theorem 3.5 relies on the following result (cf. [7]).

PROPOSITION 6.3. *If f is a holomorphic function in U^s , continuous on $[\text{cl}(U)]^s$, then*

$$(6.1) \quad f(D \cup T^s) = f([\text{cl}(U)]^s)$$

where $D := \{(z, \dots, z) : z \in \text{cl}(U)\}$.

The following corollary of Proposition 6.3 will be useful.

COROLLARY 6.4. *Let E be an open set as defined in (3.8). If P is a polynomial such that P_h is zero-free on $\text{cl}(\pi(F^s))$, then P is E -stable if and only if each of the following conditions is satisfied:*

- (a) $P(z, \dots, z)$ is zero-free on F , and
- (b) P is zero-free on $\varphi(T)^s$.

Proof of Corollary 6.4. We consider the following function:

$$f(\mathbf{z}) := \frac{P(\varphi(z_1), \dots, \varphi(z_s))}{P_h(\varphi(z_1), \dots, \varphi(z_s))}.$$

In view of the assumption on P_h , f satisfies the conditions of Proposition 6.3. Moreover, if $|(\varphi(z_1^{(w)}), \dots, \varphi(z_s^{(m)}))| \rightarrow \infty$ as $m \rightarrow \infty$ then $f(\mathbf{z}^{(m)}) \rightarrow 1$. Consequently, Corollary 6.4 is easy to deduce from Proposition 6.3. \square

Theorem 3.5 can now be established in a manner completely analogous to the proof of Theorem 4.2. We omit the details. In the proofs of the theorems in §4, the following proposition plays a crucial role.

PROPOSITION 6.5. *The perturbation function $\delta(\cdot) := \delta(\cdot; E, n, s, \|\cdot\|)$ is continuous as a function on \mathbb{C}^N (or \mathbb{R}^{2N}).*

Proof of Proposition 6.5. Since δ is an infimum of continuous functions, it is upper semicontinuous. To show that it is also lower semicontinuous, we show that for every $\lambda \geq 0$, the set $\{\mathbf{a} \in \mathbb{C}^N : \delta(\mathbf{a}) \leq \lambda\}$ is closed. Let $\mathbf{a}^{(k)} \in \mathbb{C}^N$, $\delta(\mathbf{a}^{(k)}) \leq \lambda$ for $k = 1, 2, \dots$, and $\mathbf{a}^{(k)} \rightarrow \mathbf{a}$. The proof will be complete if we can show that $\delta(\mathbf{a}) \leq \lambda$. Let $\epsilon > 0$ be arbitrary. For each $k = 1, 2, \dots$, we choose a plane $\Pi^{(k)}$ (which is either $\Pi_{\mathbf{z}}$ or $\Pi_{\mathbf{z}}^h$ for some \mathbf{z} as in (3.1a) or (3.1b)) and a vector $\mathbf{b}^{(k)} \in \Pi^{(k)}$ such that $\|\mathbf{a}^{(k)} - \mathbf{b}^{(k)}\| \leq \delta(\mathbf{a}^{(k)}) + \epsilon \leq \lambda + \epsilon$. Next, we choose an integer M such that $k \geq M$ implies $\|\mathbf{a} - \mathbf{a}^{(k)}\| < \epsilon$. Then for all $k \geq M$, we have

$$\delta(\mathbf{a}) \leq d(\mathbf{a}, \Pi^{(k)}) \leq \|\mathbf{a} - \mathbf{b}^{(k)}\| \leq \|\mathbf{a} - \mathbf{a}^{(k)}\| + \|\mathbf{a}^{(k)} - \mathbf{b}^{(k)}\| \leq \lambda + 2\epsilon.$$

Since $\epsilon > 0$ is arbitrary, this implies that $\delta(\mathbf{a}) \leq \lambda$ as desired. \square

An important role is also played by the following proposition.

PROPOSITION 6.6. *Let $E \subseteq \mathbb{C}^s$ be open, $\mathbf{a} \in \mathbb{C}^N$, E -stable, \mathbf{b} not E -stable, and Γ be any continuous path from \mathbf{a} to \mathbf{b} . Then there exists a \mathbf{c} on Γ for which $\delta(\mathbf{c}; E, n, s, \|\cdot\|) = 0$.*

The proof of Proposition 6.6 is similar to the proof of Theorem 3.4. Theorem 4.1 is a simple consequence of Proposition 6.6, Proposition 6.5, and Theorem 3.4. Also, Theorem 4.2 follows immediately from Theorem 3.4.

To prove Theorem 4.3, we apply the Hahn–Banach theorem to get

$$(6.3a) \quad d(\mathbf{a}, \Pi_{\mathbf{z}}) = \sup_{(\lambda, \mu) \in \Gamma} \frac{\lambda \operatorname{Re} P(\mathbf{a}, \mathbf{z}) + \mu \operatorname{Im} P(\mathbf{a}, \mathbf{z})}{\|(\lambda + i\mu)(\mathbf{z})\|^*};$$

$$(6.3b) \quad d(\mathbf{a}, \Pi_{\mathbf{z}}^h) = \sup_{(\lambda, \mu) \in \Gamma} \frac{\lambda \operatorname{Re} P_h(\mathbf{a}, \mathbf{z}) + \mu \operatorname{Im} P_h(\mathbf{a}, \mathbf{z})}{\|(\lambda + i\mu)(\mathbf{z}^h)\|^*},$$

where Γ is any closed curve in \mathbb{R}^2 surrounding $\mathbf{0}$ and the dual norm is the dual norm on \mathbb{R}^{2N} . Theorem 4.3 follows easily from these formulas and Proposition 6.6.

REFERENCES

- [1] B. R. BARMISH, *A generalization of Kharitonov's four-polynomial concept for robust stability with linearly dependent coefficient perturbations*, IEEE Trans. Automat. Control, 34 (1989), pp. 157–165.
- [2] N. K. BOSE, *Robust multivariate scattering Hurwitz interval polynomials*, Linear Algebra Appl., 98 (1988), pp. 123–136.
- [3] N. K. BOSE AND E. ZEHEB, *Kharitonov's theorem and stability test of multidimensional digital filters*, in Proc. IEE-G, 133 (1986), pp. 187–190.
- [4] C. K. CHUI AND H. N. MHASKAR, *A general study of maximal robust stability regions*, Circuits Systems Signal Process., 10 (1991), pp. 15–30.
- [5] K. D. KIM AND N. K. BOSE, *Invariance of the strict Hurwitz property for bivariate polynomials under coefficient perturbation*, IEEE Trans. Automat. Control, 33 (1988), pp. 1172–1174.
- [6] S. G. KRANTZ, *Function Theory of several complex variables*, John Wiley, New York, 1982.
- [7] W. RUDIN, *Function Theory in Polydiscs*, W. A. Benjamin, New York, 1969.
- [8] ———, *Functional Analysis*, McGraw-Hill, New York, 1973.

RATE OF CONVERGENCE OF RECURSIVE ESTIMATORS*

LÁSZLÓ GERENCSÉR†

Abstract. It is proved that the sequence of recursive estimators generated by Ljung's scheme combined with a suitable restarting mechanism converges under certain conditions with rate $O_M(n^{-1/2})$, where the rate is measured by the L_q -norm of the estimation error for any $1 \leq q < \infty$.

Key words. linear stochastic systems, recursive estimation, Ljung's scheme, limit theorem for moments, L -mixing processes, maximal inequalities

AMS(MOS) subject classifications. 93E12, 62F12

1. Introduction and new results. The aim of this paper is to prove a rate of convergence theorem for a class of stochastic approximation processes that has an important role in the theory of recursive identification and adaptive control of linear stochastic systems (Theorem 4.1). This class or scheme has been introduced independently by Ljung [32] and Djereveckii and Fradko [7].

The analysis presented in Ljung's paper is very complex and is not complete. Particularly, a certain "boundedness condition" imposed on the state vector is hard to verify. Also, to keep the estimator process in a compact domain, he uses a "projection mechanism" that is not adequately described and analyzed. It turns out that the analysis of the effects of the "projection" or "resetting" is one of the hardest parts of the analysis, at least in the present paper.

The work of Djereveckii and Fradko contains very useful and rigorous mathematical devices. Specially, they seem to be the first to realize the importance of moment-inequalities for the partial sums of certain kinds of mixing processes, a result due to Yoshihara [43]. An extension of Yoshihara's result to the partial sum or integral of so-called L -mixing processes as given in Gerencsér [12] is crucial in our analysis, too. A limitation of the result of Djereveckii and Fradko is that it seems they can handle systems with bounded input noise only. Also, their analysis is very complex and hard to see whether it is applicable under slightly different conditions. In spite of all these limitations it is unfortunate that their work went almost completely unnoticed in the English literature.

Further contributions to the understanding of Ljung's scheme are in Kushner and Clark [28] and Kushner [29], where the applicability of weak convergence theory for the analysis of stochastic approximation processes with "state-dependent noise" has been shown. Although the idea of looking at the evolution of "tails" of the estimator sequence or estimator process is an old one, it is carried out much further by Kushner than in previous works. His weak convergence approach has recently been simplified by Yin [42]. Another set of rigorous results have been obtained by Chen and his students (c.f. [5] for a recent survey) but their method is regression based and is not applicable to analyze the RML method in the general case. Recursive estimation

*Received by the editors June 6, 1990; accepted for publication (in revised form) May 21, 1991. This research was supported in part by Natural Sciences and Engineering Research Council grant 01329 while the author was visiting the Department of Electrical Engineering, McGill University, Montreal, Quebec, Canada, 1988-1991.

†Computer and Automation Institute of the Hungarian Academy of Sciences, H-1111, Budapest Kende u 13-17, Hungary.

within the framework of continuous-time adaptive control is considered in Duncan and Pasik-Duncan [8], [9].

Some of the practical and useful identification methods have been analysed with success using martingale techniques and/or a stochastic regression approach in Hannan [24]; Moore [34]; Solo [39], [40]; Goodwin, Ramadge, and Caines [21]; Hall and Heyde [23]; Lai and Wei [31]; Goodwin and Sin [22]; Chen [4]; Davis and Vinter [6]; and Benveniste, Metivier, and Prioruet [1].

In spite of all this progress the complexity of the analysis has not been significantly reduced and its flexibility also seems very limited. Moreover certain properties of the estimator process, such as rate of convergence of higher order moments of the estimation error, were not addressed. It turns out that this rate of convergence problem is vital to solving other more up-to-date problems, such as strong approximation of recursive ("on-line") estimators by nonrecursive ("off-line") estimators (c.f. [13]), or the almost sure asymptotics of Rissanen's predictive stochastic complexity (c.f. [14]). It should be mentioned however that Lai has recently developed a rigorous analysis of Ljung's scheme [30], but he has not considered the rate of convergence of moments issue. In another line of recent rigorous developments Heunis proved invariance principles for a class of recursive estimators, defined by the so-called Widrow algorithm [27].

Here we present a method that reveals the behaviour of the estimator process in what we think is a much better way than previous methods. The idea is similar to those applied in Ljung [32] and Djereveckii and Fradko [7], but the analysis in each step is carried out in a more careful way. The analysis we present here can be broken into roughly two parts: first we characterize stochastic approximation processes, using a random L -mixing field to update the estimator sequence. Then we relate Ljung's scheme to such stochastic approximation processes. The first step is carried out in §§1–4, while the second step is the subject of §§5–7.

As far as the technical details are concerned, the first part of the paper has certain similarities with the work of Borodin [2], who presented a continuous-time stochastic approximation scheme with a mixing field on the right-hand side. He seems to be the first author who proved a rate of convergence result, and even the weak convergence of the appropriately normalized and scaled estimation error process to a Gaussian process was proved. However he used different mixing notions and no resetting mechanism.

In contrast with his work we allow the Lipschitz-constants of the random field to be time and ω -dependent, which enables us to apply our results for linear systems with unbounded, say, Gaussian, driving noise. Another special feature of our result is that we explicitly incorporate a mechanism, which keeps our process in a prescribed "domain of stability."

The analysis we present was also partly motivated by works of Geman [11]. A novelty of our analysis compared to his is the extensive use of the theory of L -mixing processes, developed in [12].

Besides providing a rate of convergence result, (c.f. Theorem 4.1), the advantage of our approach is its flexibility, thus, for example, it enables us to derive similar results for certain continuous-time recursive estimation processes described by diffusion processes. For an earlier attempt see Gerencsér, Gyöngy, and Michaletzky [19]; for a recent independent result see Wiberg [41]. Also the analysis presented here extends to fixed gain stochastic approximation processes and is given in Gerencsér [16].

Finally a convention: in the various estimations below we shall frequently have constants which depend only on the constants that appear in the conditions below (Conditions 1.1–1.6 or later Conditions 4.1–4.2 or Conditions 4.3–4.6). These con-

stants will be called system constants. The set of real numbers will be denoted by \mathbf{R} , the k -dimensional Euclidean space will be denoted by \mathbf{R}^k . Let $D \subset \mathbf{R}^p$ be an open domain and let the stochastic process $(u_t(\theta))$ be defined on the parameter set $\mathbf{R}^+ \times D$, where $\mathbf{R}^+ = \{t : t \geq 0\}$. We begin the technical discussion with some basic definitions introduced in Gerencsér [12].

DEFINITION 1.1. We say that $(u_t(\theta))$ is M -bounded if, for all $1 \leq q < \infty$,

$$M_q(u) = \sup_{\substack{t \geq 0 \\ \theta \in D}} \mathbf{E}^{1/q} |u_t(\theta)|^q < \infty.$$

We shall use the same terminology if θ or t degenerate into a single point.

If $(u_t(\theta))$ or (u_t) is M -bounded we also write $u_t = O_M(1)$. Moreover if c_t is a sequence of positive numbers then we write $u_t = O_M(c_t)$ if $u_t/c_t = O_M(1)$.

Let a probability space (Ω, \mathcal{F}, P) be given together with a pair of families of σ -algebras $(\mathcal{F}_t, \mathcal{F}_t^+)$ such that (i) $\mathcal{F}_t \subset \mathcal{F}$ is monotone increasing, (ii) $\mathcal{F}_t^+ \subset \mathcal{F}$ is monotone decreasing and \mathcal{F}_t^+ is right continuous in t , i.e., $\mathcal{F}_s^+ = \sigma\{\bigcup_{0 < \varepsilon} \mathcal{F}_{s+\varepsilon}^+\}$, (iii) \mathcal{F}_t and \mathcal{F}_t^+ are independent for all t . For $s < 0$ we set $\mathcal{F}_s^+ = \mathcal{F}_0^+$.

DEFINITION 1.2. A stochastic process $(u_t(\theta)), t \geq 0, \theta \in D$, is L -mixing with respect to $(\mathcal{F}_t, \mathcal{F}_t^+)$ uniformly in $\theta \in D$ if it is \mathcal{F}_t progressively measurable, M -bounded, and with

$$\gamma_q(\tau, u) = \gamma_q(\tau) = \sup_{\substack{t \geq \tau \\ \theta \in D}} \mathbf{E}^{1/q} |u_t(\theta) - \mathbf{E}(u_t(\theta) | \mathcal{F}_{t-\tau}^+)|^q, \quad \tau \geq 0,$$

we have

$$\Gamma_q = \Gamma_q(u) = \int_0^\infty \gamma_q(\tau) d\tau < \infty.$$

The definition extends to parameter-free processes (u_t) and to discrete-time processes in an obvious way.

Define the process $\Delta u / \Delta \theta$ by $\Delta u / \Delta \theta = |u_n(\theta + h) - u_n(\theta)| / |h|$ defined for $n \geq 0, \theta \neq \theta + h \in D$.

DEFINITION 1.3. The stochastic process $u_n(\theta)$ is M -Lipschitz-continuous in θ if the process $\Delta u / \Delta \theta$ is M -bounded, i.e., if for all $1 \leq q < \infty$, we have

$$M_q(\Delta u / \Delta \theta) = \sup_{\substack{n \geq 0 \\ \theta \neq \theta + h \in D}} \mathbf{E}^{1/q} |u_n(\theta + h) - u_n(\theta)|^q / |h| < \infty.$$

The main object of our study is a random differential equation of the form

$$(1.1) \quad \dot{x}_t = \frac{1}{t} H(t, x_t, \omega),$$

where $H = (H(t, x, \omega))$ is a random field defined in $[1, \infty) \times D$, where D is a bounded open domain in $\mathbf{R}^p \times \Omega$. Define another random field $\Delta H / \Delta x$ by

$$\Delta H / \Delta x(t, x, x + h, \omega) = |H(t, x + h, \omega) - H(t, x, \omega)| / |h|$$

where $x, x + h \in D, h \neq 0$. Here $|\cdot|$ denotes the Euclidean norm.

CONDITION 1.1. The processes $(H(t, x, \omega))$ and $(\Delta H / \Delta x(t, x, x + h, \omega))$ are separable and L -mixing with respect to $(\mathcal{F}_t, \mathcal{F}_t^+)$ uniformly in $x, x + h \in D$.

DEFINITION 1.4. Let $(u_t), t \geq 0$ be a real-valued stochastic process. We say that (u_t) is in class M^* if for some $\varepsilon > 0$

$$M^\varepsilon(x) \triangleq \sup_t \frac{1}{\varepsilon} \log E \exp \varepsilon u_t < \infty.$$

The definition extends to vector-valued and parameter-dependent processes say $x_t(\theta)$ in a natural way. In the latter case we take the supremum over t and θ .

CONDITION 1.2. $H(t, x, \omega)$ is continuous in t and it is Lipschitz-continuous in x with a (t, ω) -dependent Lipschitz constant $L_t = L_t(\omega)$, i.e.,

$$|H(t, x, \omega) - H(t, x', \omega)| \leq L_t(\omega)|x - x'|$$

where $L_t(\omega)$ is in class M^* .

CONDITION 1.3. We have

$$EH(t, x, \omega) = G(x) + \delta \tilde{G}(t, x),$$

where $\delta \tilde{G}(t, x) = O(t^{-1/2})$ uniformly in x , and $G(x)$ satisfies the conditions below.

Let the ordinary differential equation

$$(1.2) \quad \dot{y}_t = \frac{1}{t} G(y_t), \quad y_s = \xi,$$

$t \geq s$ satisfy the following conditions.

CONDITION 1.4. $G(y)$ is defined in D and it has continuous and bounded partial derivatives up to second order, say, $\|\partial G / \partial y\| < L$ and $\|\partial^2 G / \partial y^2\| < L$. (Here $\|\cdot\|$ denotes the operator norm of a matrix.) Also assume that $G(0) = 0$.

Under the condition above (1.2) has a unique solution in $[s, \infty)$, which we denote by $y(t, s, \xi)$. It is well known (c.f. Pontryagin [35, Chap. 24, Thm. 17]) or Hartman [26, Chap. V, Thm. 1.1]) that $y(t, s, \xi)$ is a continuously differentiable function of (t, s, ξ) . We impose a stability condition onto (1.2) to be formulated below. Let us introduce the notation $\psi(t, s, \xi) = (\partial / \partial \xi) y(t, s, \xi)$ for $s \leq t$. Let $D_0 \subset \text{int} D$ denote a compact domain such that $0 \in \text{int} D_0$. Then $\psi(t, s, \xi), s \leq t$ is a $p \times p$ matrix-valued function that solves the equation in variations

$$\frac{d}{dt} \psi(t, s, \xi) = \frac{1}{t} \frac{\partial}{\partial y} G(y_t) \cdot \psi(t, s, \xi), \quad \psi(s, s, \xi) = I.$$

CONDITION 1.5. For every $\xi \in D_0$, $t > s > 0$ $y(t, s, \xi) \in D$ is defined and we have with some $C_0, \alpha > 0$,

$$(1.3) \quad \|(\partial / \partial \xi) y(t, s, \xi)\| = \|\psi(t, s, \xi)\| \leq C_0(s/t)^\alpha.$$

Furthermore we assume that the initial condition $\xi \in \text{int} D_{00} \subset \text{int} D_0$, where D_{00} is a compact domain which is invariant for (1.2), and that for any $t > s > 0$,

$$y(t, s, D_{00}) = \{y(t, s, x) : x \in D_{00}\} \subset \text{int} D_{00}.$$

The inequality (1.3) is equivalent to the condition that the differential equation

$$(1.2)' \quad \frac{d}{dv} z_v = G(z_v), \quad z_u = \xi$$

is exponentially asymptotically stable with exponent α , i.e., if the solutions of (1.2)' are denoted by $z(v, u, \xi)$, then we have

$$\|(\partial/\partial\xi)z(v, u, \xi)\| \leq C_0 e^{-\alpha(u-v)}.$$

This is obtained by a simple change of time-scale $t = e^v$, $s = e^u$.

Let us assume that our actual data is $H(t, x, \omega) + \delta\tilde{H}(t, \omega)$ and let us consider the random differential equation

$$(1.4) \quad \dot{x}_t = \frac{1}{t}(H(t, x_t, \omega) + \delta\tilde{H}(t, \omega)), \quad x_1 = \xi,$$

with $\xi \in \text{int}D_{00}$. When x_t hits the boundary ∂D_0 of D_0 , say at τ , we set $x_{\tau+} = \xi$. We shall see that we get a piecewise continuous trajectory x_t defined for $t \in [1, \infty)$.

CONDITION 1.6. $(\delta\tilde{H}(t, \omega))$ is a measurable process, continuous in t almost surely. Moreover if $\tau(\sigma)$ denotes the first moment after σ when x_t hits ∂D_0 and $q > 1$ is a fixed real number, then

$$(1.5) \quad \sup_{s \leq \sigma \leq qs} \int_{\sigma}^{\tau(\sigma) \wedge q\sigma} \frac{1}{r} |\delta\tilde{H}(r, \omega)| dr = O_M(s^{-1/2}).$$

Remark. Note that (1.5) is obviously satisfied if $|\delta\tilde{H}(r, \omega)| = O_M(r^{-1/2})$, due to the triangle inequality for $L_q(\Omega, \mathcal{F}, P)$ -norms, since the measure dr/r integrates to $\log(q-1)$ in $[s, qs)$. However in our application such a priori bound cannot be derived.

THEOREM 1.1. *Under Conditions 1.1–1.6, x_t is defined for all $t \in [1, \infty)$ and if $\alpha > 1/2$ then $x_t = O_M(t^{-1/2})$. Moreover the following stronger result also holds: let $1 < q < \infty$ and*

$$x_n^* = \sup_{q^n \leq t < q^{n+1}} |x_t|,$$

then $x_n^ = O_M(n^{-1/2})$. For $\alpha < 1/2$ we have for any $\varepsilon > 0$, $x_n^* = O_M(n^{-\alpha+\varepsilon})$.*

Some important corollaries of this theorem will be described in §5 (c.f. Theorem 4.2 and its corollaries).

Let us now consider a discrete-time process

$$x_{n+1} = x_n + \frac{1}{n+1}(H(n, x_n, \omega) + \delta\tilde{H}(n, \omega)), \quad x_0 = \xi \in \text{int } D_0.$$

Let $(H^c(t, x, \omega))$ and $(\delta\tilde{H}^c(t, \omega))$ be the piecewise constant continuous-time extensions of $(H(n, x, \omega))$ and $(\delta\tilde{H}(n, \omega))$, respectively. That is, $H^c(t, x, \omega) = H(n, x, \omega)$ for $n \leq t < n+1$ and $\delta\tilde{H}^c(t, \omega) = \delta\tilde{H}(n, \omega)$ for $n \leq t < n+1$. If x_{n+1} leaves D_0 , then we redefine x_{n+1} to be x_0 .

THEOREM 1.2. *Assume that the random field $(H^c(t, x, \omega))$ and the stochastic process $\delta\tilde{H}^c(t, \omega)$ satisfy Conditions 1.1–1.6, except that continuity in t is replaced by piecewise continuity with possible jumps at integers. Then*

$$x_n = O_M(n^{-1/2}).$$

2. The proofs of Theorems 1.1 and 1.2. For the sake of simplicity assume that $\delta\tilde{G}_t = \delta\tilde{H}(t, \omega) = 0$ for all t . The general case will be considered at the end of the section. Let $s \geq 1$ and $s \leq \sigma < qs$. Let $\tau(\sigma)$ denote the first moment after σ , at which

x_t hits ∂D_0 . Further, let \bar{y}_t denote the solution of (1.2) with initial condition $\bar{y}_\sigma = x_\sigma$, i.e., $\bar{y}_t = y(t, \sigma, x_\sigma)$. A main technical tool used in the proof is the development of an upper bound for the increments $|x_t - \bar{y}_t|$. Let

$$I_\sigma(q) = \sup_{\sigma \leq t < \tau(\sigma) \wedge q\sigma} |x_t - \bar{y}_t|$$

and

$$I_s^*(q) = \sup_{s \leq \sigma \leq qs} I_\sigma(q)$$

LEMMA 2.1. We have $I_s^*(q) = O_M(s^{-1/2})$.

Proof. Let

$$\bar{H}(t, x, \omega) = H(t, x, \omega) - G(x),$$

then we have for $\sigma \leq t < \tau(\sigma)$,

$$\begin{aligned} |x_t - \bar{y}_t| &= \left| \int_\sigma^t \frac{1}{r} (H(r, x_r, \omega) - G(\bar{y}_r)) dr \right| \\ &= \left| \int_\sigma^t \frac{1}{r} (\bar{H}(r, x_r, \omega) + G(x_r) - G(\bar{y}_r)) dr \right|. \end{aligned}$$

Let us write $\bar{H}(r, x_r, \omega) = \bar{H}(r, \bar{y}_r, \omega) + \bar{H}(r, x_r, \omega) - \bar{H}(r, \bar{y}_r, \omega)$. Taking into account the Lipschitz-condition imposed on H (Condition 1.2) we get after easy calculations

$$(2.1) \quad |x_t - \bar{y}_t| \leq \int_\sigma^t \frac{1}{r} (L_r + L) |x_r - \bar{y}_r| dr + \left| \int_\sigma^t \frac{1}{r} \bar{H}(r, \bar{y}_r, \omega) dr \right|.$$

Let us consider the expression

$$(2.2) \quad \delta_s^*(q) = \sup_{\substack{s \leq \sigma \leq qs \\ \sigma \leq t \leq q\sigma \\ x \in D_0}} \left| \int_\sigma^t \frac{1}{r} \bar{H}(r, y(r, \sigma, x), \omega) dr \right|,$$

where the supremum is taken over σ , t , and x . Let $u_r(\sigma, x) = \bar{H}(r, y(r, \sigma, x), \omega)$, $r \geq \sigma$. Then we can write

$$(2.3) \quad \delta_s^*(q) = \sup_{\substack{s \leq \sigma \leq qs \\ \sigma \leq t \leq q\sigma \\ x \in D_0}} \left| \int_\sigma^t \frac{1}{r} u_r(\sigma, x) dr \right|.$$

It will be shown in the next section that $\delta_s^*(q) = O_M(s^{-1/2})$ (c.f. Lemma 3.1).

Let us now set $z_t = t^\alpha |x_t - \bar{y}_t|$. Multiplying (2.1) by t^α we get for $\sigma \leq t < \tau(\sigma)$,

$$z_t \leq t^\alpha \delta_s^*(q) + \int_\sigma^t t^\alpha \frac{1}{r} (L_r + L) \frac{z_r}{r^\alpha} dr \leq (q\sigma)^\alpha \delta_s^*(q) + \int_\sigma^t \frac{C}{r} (L_r + L) z_r dr,$$

with $C = q^\alpha$. From this the Bellman–Gronwall lemma gives $z_t \leq \kappa_s(q)(q\sigma)^\alpha \delta_s^*(q)$ with

$$(2.4) \quad \kappa_s(q) = \exp \int_s^{q^2 s} \frac{C}{r} (L_r + L) dr,$$

and we get

$$(2.5) \quad |x_t - \bar{y}_t| = t^{-\alpha} z_t \leq C' \kappa_s(q) \delta_s^*(q)$$

where $C' = q^{2\alpha}$.

We show that $\kappa_s(q)$ has finite moment of order m for any given m whenever $q - 1$ is sufficiently small. Let $\bar{q} = C \log q^2$ and apply Jensen's inequality with the measure $(C/r)(dr/\bar{q})$ to get

$$(\kappa_s(q))^m = \exp \left\{ m\bar{q} \int_s^{q^2 s} (L_r + L) \frac{C}{r} \frac{dr}{\bar{q}} \right\} \leq \int_\sigma^{q^2 s} \exp \{ m\bar{q}(L_r + L) \} \frac{C}{r} \frac{dr}{\bar{q}}.$$

Taking expectation of both sides and noticing that $L_r(\omega)$ is in class M^* we get that $E(\kappa_s(q))^m < \infty$, whenever $m\bar{q} < \varepsilon$. Since $\bar{q} = 2C \log q < 2C(q - 1)$, the last condition is certainly satisfied if $q - 1 < \varepsilon/2Cm$, where $\varepsilon > 0$ is the same small positive number that had been introduced in the definition of the class M^* (c.f. Definition 1.4). Combining the above estimates we get

$$I_s^*(q) \leq C \kappa_s(q) \delta_s^*(q)$$

and here $E|\kappa_s(q)|^m < \infty$ if $(q - 1) < \varepsilon/2Cm$.

Let now $q > 1$ be arbitrary and let us subdivide the interval $[s, qs]$ into a fixed number of subintervals as follows: define $q' = q^{1/m'}$, where m' is a natural number and the i th point of the subdivision is defined as $s_i = s(q')^i$, $i = 1, \dots, m'$. Similarly for any σ such that $s \leq \sigma \leq qs$, define $\sigma_i = \sigma(q')^i$, $i = 1, \dots, m'$. Then we have

$$(2.6) \quad I_{\sigma_i}(q') \leq I_{\sigma_i}^*(q') \leq C \kappa_{\sigma_i}(q') \delta_{\sigma_i}^*(q'),$$

and here $E|\kappa_{\sigma_i}(q')|^m < \infty$ for $m < \varepsilon/2C(q' - 1)$. Since m' can be chosen arbitrarily, we can make $(q' - 1)$ as small as we like and hence make m as large as we like.

Now, the stability condition (Condition 1.5) implies that in the interval $\sigma \leq t \leq \tau(\sigma) \wedge qs$ we have

$$(2.7) \quad |x_t - \bar{y}_t| \leq C_0 \sum_{i=0}^{N_t-1} (q')^{-\alpha(N_t-i-1)} I_{\sigma_i}(q'),$$

where N_t is the index of that interval in the subdivision that contains t . Hence in $\sigma \leq t \leq \tau(\sigma) \wedge qs$ we certainly have

$$|x_t - \bar{y}_t| \leq C_0 \sum_{i=0}^{m'-1} I_{\sigma_i}(q').$$

On the other hand, for any σ such that $s \leq \sigma \leq qs$ we have $I_\sigma(q') \leq I_{s_j}^*(q')$, where s_j is such that $s_j \leq \sigma \leq s_{j+1}$. Hence we finally get that for any $\sigma \leq t \leq \tau(\sigma) \wedge qs$

$$(2.8) \quad |x_t - \bar{y}_t| \leq C_0 \sum_{i=0}^{m'-1} I_{s_j}^*(q'),$$

and here the right-hand side is independent of σ and t , and thus it is an upper bound for $I_s^*(q)$.

Now for a given m choose q' so that $(q')^2 - 1 < \varepsilon/2Cm$. Then the right-hand side is the sum of a fixed number of terms each being in $L_m(\Omega, \mathcal{F}, P)$; moreover,

$$E^{1/m}(I_{s_j}^{*m}(q')) = O_M(s^{-1/2}),$$

and hence the proposition of the lemma follows. \square

LEMMA 2.2. x_t is defined in the whole interval $[s, qs]$ with probability 1.

Proof. Let $\tau_0 = s$ and τ_i be successive moments when x_t hits ∂D_0 . We set $\tau_i = qs$ if x_t does not hit ∂D_0 after the time $\tau_i - 1$. Let $\tau^* = \lim_{i \rightarrow \infty} \tau_i$. Then $|x_{\tau_i-} - x_{\tau_{i-1}}| > C > 0$ with some fixed $C > 0$ as long as $\tau_i < qs$. (Here x_{t-} denotes the left side limit of (x_t) at t .) Let

$$A_i = \left\{ \omega : \left| \int_{\tau_{i-1}}^{\tau_i-} \frac{1}{t} H(t, x_t, \omega) dt \right| > C \right\}$$

and

$$H^*(t, \omega) = \sup_{x \in D_0} |H(t, x, \omega)|.$$

Then we certainly have $P(A_i) \leq P(A'_i)$ where

$$A'_i = \left\{ \omega : \int_{\tau_{i-1}}^{\tau_i-} H^*(t, \omega) dt > C \right\}.$$

Furthermore

$$P(A'_i) \leq E \int_{\tau_{i-1}}^{\tau_i-} H^*(t, \omega) dt / C,$$

hence

$$\sum_{i=1}^{\infty} P(A'_i) \leq E \int_s^{qs} H^*(t, \omega) dt / C < \infty,$$

since $(H^*(t, \omega))$ is an M -bounded process by Theorem 7.2 in Appendix II. Thus it follows that only a finite number of the events A_i occur almost surely, and thus x_t is well defined in $[s, qs]$ as stated in the lemma. \square

LEMMA 2.3. Let C_s denote the event x_t hits ∂D_0 in the interval $[s, qs]$. Then we have for any $m \geq 1$

$$P(C_s) = O(s^{-m}).$$

Proof. It is sufficient to prove the lemma for some integer power of q . This implies that we can assume q as large as we wish. Assume $s \geq q$ and that x_t hits ∂D_0 in (s, qs) . Let us consider the past of the process x_t in the interval $[s/q, s]$.

If x_t hits ∂D_0 in the interval $[s/q, s]$ then consider the interval $[\tau(s/q), \tau^2(s/q)]$ where $\tau^2(\sigma) = \tau(\tau(\sigma))$. We certainly have $\tau^2(s/q) \leq \tau(s) \leq qs$. Now since x_t is reset to ξ at $\tau(s/q)$ and since the invariance of D_{00} with respect to the ordinary differential equation (1.2) implies that $\bar{y}_t = \bar{y}(t, \tau(s/q), D_{00}) \subset D_{00}$ for all t , we shall have for $t = \tau^2(s/q)$, $|x_t - \bar{y}_t| > c > 0$. But then we have $I_{s/q}^*(q^2) > c > 0$, the probability of which is $O(s^{-m})$ with any $m \geq 1$, due to Tchebishev's inequality.

Now if x_t does not hit ∂D_0 in the interval $[s/q, s]$, then we have to compare x_t with the trajectory $\bar{y}_t = y(t, s/q, x_{s/q})$ in the interval $[s/q, \tau(s)]$. If q is sufficiently large then $y(t, s/q, D_0) \subset D_{00}$ for $t > s$, hence for $t = \tau(s)$ we shall have $|x_t - \bar{y}_t| > c > 0$, so again we conclude that $I_{s/q}^*(q^2) > c > 0$, the probability of which is $O(s^{-m})$. Thus Lemma 2.3 has been proved. \square

Remark. Note that for the proof of Lemma 2.3 we used less than Lemma 2.1, namely we used only the following estimations:

$$I_{s/q}^*(q^2) = O_M(s^{-1/2}) \quad \text{and} \quad I_{\tau(s)}^*(q^2) = O_M(s^{-1/2}).$$

Combining Lemmas 2.1 and 2.3 we get the following result which is similar in spirit to Lemma 2.1.

LEMMA 2.4. *Let K denote the diameter of D_0 . Then we have*

$$(2.9) \quad \sup_{s \leq t \leq qs} |x_t - \bar{y}_t| \leq I_s^*(q) + K\chi_{C_s} = O_M(s^{-1/2}).$$

Proof of Theorem 1.1. Let $1 < q < 2$ and subdivide $[1, \infty)$ by the points $q^i, i = 0, 1, \dots$. For $1 \leq i \leq N-1$ we consider the interval $[q^i, q^{i+1})$ and approximate x_t there by the trajectory $\bar{y}_i = (\bar{y}_{it})$ which is the solution of (1.2) with initial condition $\bar{y}_{iq^i} = x_{q^i}$. Set

$$d_i^* = \sup_{q^i \leq s < q^{i+1}} |x_s - \bar{y}_{is}|.$$

Obviously $d_i^* = O_M(q^{-i/2})$.

Now the stability condition (Condition 1.5) implies that

$$(2.10) \quad |x_{q^N} - y_{q^N}| \leq C_0 \sum_{i=0}^{N-1} (q^{i+1}/q^N)^\alpha d_i^* \stackrel{\nabla}{=} r_N.$$

Since $\alpha > 1/2$ we can apply Lemma 7.4 of Appendix II, to get that

$$(2.11) \quad E^{1/m} |r_N|^m \leq C' q^{-N/2}$$

with some constant C' , which is independent of N .

For $q^N < t < q^{N+1}$ the same argument applies, with minor modification, namely, using the stability condition in the interval $[q^N, q^{N+1})$ we get

$$\sup_{q^N \leq t < q^{N+1}} |x_t - y_t| \leq C_0 r_N + d_N^* = O(q^{-N/2})$$

and thus the theorem is proved for $\alpha > 1/2$.

If the condition $\alpha > 1/2$ is relaxed to $\alpha > 0$ then (2.11) is weakened to the following: we have for any $\varepsilon > 0$

$$(2.12) \quad E^{1/m} |r_N|^m < C' q^{-\alpha+\varepsilon}.$$

The proof can be completed as for the $\alpha > 1/2$ case using the second part of Lemma 7.4.

The proof of the general case $\delta \tilde{G}(t, x) \neq 0, \delta \tilde{H}(t, \omega) \neq 0$ can easily be obtained. Indeed (2.1) has to be modified by adding the term

$$\delta I_\sigma(q) = \sup_{\sigma \leq t < q(\sigma) \wedge \tau(\sigma)} \int_\sigma^t \frac{1}{r} (|\delta \tilde{H}(r, \omega)| + |\delta \tilde{G}(r, \bar{y}_r)|) dr$$

to the right-hand side. Conditions 1.3 and 1.6 imply that

$$\delta I_s^*(q) = \sup_{s \leq \sigma \leq qs} \delta I_\sigma(q) = O_M(s^{-1/2}),$$

and the rest of the proof is unaffected. \square

Proof of Theorem 1.2. Consider the piecewise linear curve x_t^c defined as $x_t^c = (t - n)x_{n+1} + (n + 1 - t)x_n$ for $n \leq t \leq n + 1$ if $x_{n+1} \in D_0$, and $x_t^c = x_n$ for $n \leq t < n + 1$ if $x_{n+1} \notin D_0$, and in this latter case we set $x_{n+1}^c = x_0$. x_t^c can be considered as the solution of a differential equation with an anticipating resetting mechanism: $\dot{x}_t^c = x_n$ if $x_{n+1} \notin D_0$. This differential equation is obtained as follows: If x_{n+1} is not reset and x_t^c is any point constructed above, then we find a second continuous time “correction term” $\delta\tilde{H}_2(t, \omega)$ from the equality

$$\frac{1}{t}(H^c(t, x_t^c, \omega) + \delta\tilde{H}_1^c(t, \omega) + \delta\tilde{H}_2^c(t, \omega)) = \frac{1}{n+1}(H(n, x_n, \omega) + \delta\tilde{H}(n, \omega)),$$

from which we get for $\delta\tilde{H}^c(t, \omega) = \delta\tilde{H}_1^c(t, \omega) + \delta\tilde{H}_2^c(t, \omega)$ the inequality (2.13)

$$\begin{aligned} |\delta\tilde{H}^c(t, \omega)| &= \left| \frac{1}{n+1}(H(n, x_n, \omega) + \delta\tilde{H}(n, \omega)) - H^c(t, x_t^c, \omega) \right| \\ &\leq \frac{t}{n+1}L_n|x_n - x_t^c| + \left| \left(\frac{t}{n+1} - 1 \right) H^c(t, x_t^c, \omega) \right| + \left| \frac{t}{n+1} \delta\tilde{H}(n, \omega) \right|. \end{aligned}$$

The contribution of the first two terms is $O_M(n^{-1})$. Indeed, using the maximal inequality of Appendix II (Theorem 7.2) to estimate $|x_n - x_t^c|$ and $|H^c(t, x_t^c, \omega)|$ we get (2.13). This estimate and the condition imposed on $\delta\tilde{H}(n, \omega)$ imply that $\delta\tilde{H}^c(t, \omega)$ satisfies Condition 1.6 except that the resetting mechanism is now defined in a different way. It is easy to see that the proof of Theorem 1.1 is not effected by the above minor change in the restarting mechanism, thus Theorem 1.2 follows. \square

3. Taking supremum over σ . The purpose of this section is to prove the following lemma.

LEMMA 3.1. *Let us define*

$$\delta_s^*(q) = \sup_{\substack{s \leq \sigma \leq qs \\ \sigma \leq t \leq q\sigma \\ x \in D_0}} \left| \int_{\sigma}^t \frac{1}{r} \bar{H}(r, y(r, \sigma, x), \omega) dr \right|.$$

Then $\delta_s^*(q) = O_M(s^{-1/2})$.

Remark. If σ is fixed, say $\sigma = s$, then the claim of the lemma immediately follows from Theorems 7.1 and 7.2 of Appendix II. Indeed, the processes $(u_r(\sigma, x))$ and $\Delta u_r(\sigma, x)/\Delta x$ are L -mixing, uniformly in x for $x \in D_0$, and in $x, x + h \in D_0$, respectively, with respect to $(\mathcal{F}_t, \mathcal{F}_t^+)$, by Lemma 7.3 of Appendix II. They are also obviously of zero-mean.

Defining

$$\delta_{\sigma}(q) = \sup_{\substack{\sigma \leq t \leq q\sigma \\ x \in D_0}} \left| \int_{\sigma}^t \frac{1}{r} u_r(\sigma, x) dr \right|$$

we get that for any $m > 2$

$$E^{1/m} |\delta_{\sigma}(q)|^m \leq C' \left(\int_{\sigma}^{q\sigma} \frac{1}{r^2} dr \right)^{1/2} = C' \sigma^{-1/2},$$

where C' depends only on m , and on $M_{2m}(u)$ and $\Gamma_{2m}(u)$ and of the regions D_0 and D , but it is independent of q in the region $1 < q < q_0$, for any fixed q_0 .

The main difficulty of the proof is therefore handling supremum over t over a set of dilating intervals. This problem will be solved with the application of an appropriate change of time-scale.

Let us define the process $\delta_\sigma(x) = \delta_\sigma(x, q)$ by

$$\delta_\sigma(x, q) = \sup_{\sigma \leq t < q\sigma} \left| \int_\sigma^t \frac{1}{r} \overline{H}(r, y(r, \sigma, x), \omega) dr \right|$$

for $x \in D_0$. Note that the integral expression appearing in the definition of $\delta_\sigma(x)$ is identical with the one appearing in the definition of $\delta_s^*(q)$, but in the definition of $\delta_\sigma(x, q)$ we take supremum over t only. Let us set $\sigma = e^v$ and

$$\rho_v(x) = e^{v/2} \delta_{e^v}(x).$$

With this notation we have the following lemma.

LEMMA 3.2. *The processes $\rho_v(x)$, $|\rho_v(x+h) - \rho_v(x)|/|h|$, $h \neq 0$, and $|\rho_{v+k}(x) - \rho_v(x)|/|k|$, $k \neq 0$, are M -bounded.*

Proof. We have already shown that $\delta_\sigma(x) = O_M(\sigma^{-1/2})$ therefore $\rho_v(x) = O_M(1)$. The process $(|\rho_v(x+h) - \rho_v(x)|/|h|)$ can also be shown to be M -bounded, with the same method.

Let us now take a small $h > 0$ and estimate the moments of $\rho_{v+k} - \rho_v$, or equivalently, the moments of $\delta_{\sigma(1+h)} - \delta_\sigma$ where $1+h = e^k$. Note that

$$(3.1) \quad \sup_{x \in D_0} \sup_{\substack{1 \leq \sigma \leq r \\ h > 0}} |y(r, \sigma(1+h), x) - y(r, \sigma, x)|/h < \infty.$$

Indeed we have

$$\frac{\partial y}{\partial \sigma}(r, \sigma, x) = -\frac{1}{\sigma} G(x) \frac{\partial y}{\partial x}(r, \sigma, x),$$

hence

$$\sup_{x \in D_0} \sup_{1 \leq \sigma \leq r} \left| \sigma \frac{\partial y}{\partial \sigma}(r, \sigma, x) \right| \leq \sup_{x \in D_0} |G(x)| \cdot C_0 < \infty,$$

from which (3.1) follows if we expand the difference in question into a first-order Taylor series.

The difference $\delta_{\sigma(1+h)} - \delta_\sigma$ can obviously be majorated by the sum of the following three terms:

$$(3.2) \quad \Delta_1 = \sup_{\sigma \leq t \leq \sigma(1+h)} \int_\sigma^t \frac{1}{r} \overline{H}(r, y(r, \sigma, x), \omega) dr / h,$$

$$(3.3) \quad \Delta_2 = \sup_{q\sigma \leq t \leq q\sigma(1+h)} \left| \int_{q\sigma}^t \frac{1}{r} \overline{H}(r, y(r, \sigma(1+h), x), \omega) dr \right| / h,$$

and

$$(3.4) \quad \Delta_3 = \sup_{\sigma(1+h) \leq t \leq q\sigma} \left| \int_{q\sigma(1+h)}^t \frac{1}{r} \Delta^* \overline{H}_r dr \right| / h,$$

where

$$(3.5) \quad \Delta^* \bar{H}_r = (\bar{H}(r, y(r, \sigma(1+h), x), \omega) - \bar{H}(r, y(r, \sigma, x), \omega))/h.$$

It is easy to see that Δ_1 is M -bounded with respect to (σ, h) . Indeed

$$\Delta_1 \leq \int_{\sigma}^{\sigma(1+h)} \frac{1}{r} \bar{H}^*(r, \omega) dr / h$$

where

$$\bar{H}^*(r, \omega) = \sup_{x \in D_0} |\bar{H}(r, x, \omega)|$$

is M -bounded by Theorem 7.2 of Appendix II. The second term, i.e., Δ_2 , is estimated similarly.

Finally for the third term we note that

$$\Delta \bar{H}(r, y_1, y_2, \omega) = (\bar{H}(r, y_1, \omega) - \bar{H}(r, y_2, \omega)) / |y_1 - y_2|$$

is an L -mixing process, uniformly in y_1, y_2 for $y_1 \neq y_2$ in D , with respect to $(\mathcal{F}_t, \mathcal{F}_t^+)$, hence we conclude using (3.1) that $\Delta^* \bar{H}_r$ defined under (3.5) is an L -mixing process. Since $\Delta \bar{H}_r$ has zero-mean we can apply once more the moment inequality given as Theorem 7.1 to conclude that Δ_3 is M -bounded. \square

LEMMA 3.3. *Let $h > 0$ be fixed and define the process*

$$\rho_v^* = \sup_{\substack{v \leq z \leq v+h \\ x \in D_0}} \rho_z.$$

Then the process (ρ_v^) is M -bounded.*

Proof. The proof is obtained by applying the maximal inequality given as Theorem 7.2 of Appendix II to the congruent compact domains $[v, v+h] \times D_0$ with varying v . \square

Proof of Lemma 3.1. Observing that $\delta_s^*(q) = s^{-1/2} \rho_w^*$ under the correspondence $s = e^w$, we immediately get the proposition. \square

4. Rate of convergence for Ljung's scheme. An important observation in the theory of recursive identification and adaptive control of linear stochastic systems is that most of the ad hoc procedures can be described by a general scheme. This general structure was discovered independently by Ljung on the one hand and Djereveckii and Fradko on the other hand (c.f. [32] and [7].) This general setup and the associated recursive estimation method is sometimes called Ljung's scheme. The applicability of Ljung's scheme in recursive identification has been demonstrated in Ljung and Söderström [33]. Also it has recently been shown in Gerencsér [15] that a wide class of stochastic adaptive control problems can be solved using Ljung's scheme.

The estimation problem itself can be formulated as follows: First we consider a state-space equation of the form

$$(4.1) \quad \bar{\phi}_{n+1}(x) = A(x) \bar{\phi}_n(x) + B(x) e_n, \quad \bar{\phi}_0(x) = 0,$$

with $x \in D \subset \mathbb{R}^p$, $\phi(x) \in \mathbb{R}^r$, and $e_n \in \mathbb{R}^m$. We assume the following conditions.

CONDITION 4.1. The family of $r \times r$ matrices $A(x)$, $x \in D \subset \mathbb{R}^p$ are jointly stable in the following sense: there exists a positive-definite $n \times n$ matrix V such that

$$A^T(x) V A(x) \leq \lambda V \quad \text{with some } 0 < \lambda < 1$$

for all $x \in D$. Moreover the functions $(A(x)), (B(x))$ are twice continuously differentiable with bounded partial derivatives up to second order in D_0 .

Let us remember that we defined for any positive integer τ and $q \geq 1$

$$\gamma_q(\tau, e) = \sup_{n \geq \tau} E^{1/q} |e_n - (e_n | \mathcal{F}_{n-\tau}^+)|^q.$$

This quantity is essential in the definition of L -mixing processes. It has been shown in Gerencsér [14] that if (u_n) is an L -mixing process then $\gamma_q(\tau, u) \leq 16\Gamma_q(u)/\tau$ for all $q \geq 1$ and $\tau \geq 1$ integers. We shall need a slightly stronger condition to be satisfied by (e_n) .

DEFINITION 4.1. We say that a stochastic process is L^+ -mixing if for all integer $\tau \geq 1$ and $q \geq 1$ with some $c > 0$

$$\gamma_q(\tau, e) = O(\tau^{-1-c})$$

CONDITION 4.2. We assume that (e_n) is wide sense stationary (e_n^2) is in class M^* and it is L^+ -mixing with respect to a pair of families of σ -algebras $(\mathcal{F}_n, \mathcal{F}_n^+)$.

Let Q be a quadratic function from \mathbf{R}^r to \mathbf{R}^p . Define

$$G(x) = \lim_{n \rightarrow \infty} EQ(\bar{\phi}_n(x)).$$

It is easy to see that $G(x)$ is well defined in D and has continuous and bounded partial derivatives up to second order, say, $\|\partial G/\partial y\| \leq L$ and $\|\partial^2 G/\partial y^2\| \leq L$. (Here $\|\cdot\|$ denotes the operator norm of a matrix.). The problem can now be formulated as follows: solve the nonlinear algebraic equation

$$G(x) = 0.$$

We shall assume that the solution is at $x^* = 0$.

Let $x_0 = \xi$ be an initial estimate of x^* . Then the description of Ljung's scheme is completed by the following recursion:

$$(4.3) \quad \phi_{n+1} = A(x_n)\phi_n + B(x_n)e_n, \quad \phi_0 = 0.$$

$$(4.2) \quad x_{n+1} = x_n + \frac{1}{n+1}Q(\phi_{n+1})$$

We shall have to keep (x_n) in a compact domain D_0 and therefore if x_{n+1} leaves D_0 we redefine it to be x_0 . To formalize this procedure let x_{n+1-} denote the value of x computed at time $n+1$ by (4.2) and let x_{n+1} be the actual value, which is x_0 if $x_{n+1-} \in D_0^c$. Let $B_{n+1} = \{\omega : x_{n+1-} \in D_0^c\}$, then (4.3) will be replaced by

$$(4.4) \quad x_{n+1} = x_n + (1 - \chi_{B_{n+1}})\frac{1}{n+1}Q(\phi_{n+1}) + \chi_{B_{n+1}}(x_0 - x_n).$$

As we shall see the analysis of the effect of resetting in the context of Ljung's scheme is technically not easy.

Let us consider the so-called associated differential equation

$$(4.5) \quad \dot{y}_t = \frac{1}{t}G(y_t), \quad y_s = \xi.$$

Under the condition above (4.5) has a unique solution in some interval, which we denote by $y(t, s, \xi)$, and it is a continuously differentiable function of (t, s, ξ) .

A main result of the paper is the following theorem, which improves Ljung [32].

THEOREM 4.1. *Assume that the differential equation (4.5) satisfies Condition 1.5 with $\alpha > 1/2$, and Conditions 4.1–4.2 are satisfied. Then we have $x_n = O_M(n^{-1/2})$.*

The theorem implies almost sure convergence of x_n . Indeed $P(|x_n| > c > 0) = O(n^{-m})$ for any $m \geq 1$ by Theorem 4.1 and Chebyshev's inequality, hence a Borel–Cantelli argument yields $x_n \rightarrow 0$ almost surely.

As an application of Theorem 4.1 we shall state a theorem on the rate of convergence of a recursive prediction error estimator for ARMA processes. We say “a recursive prediction error estimator” rather than “the recursive prediction error estimator” since there are many similar methods all of which are called recursive prediction error estimation methods. The special feature of the method we propose here is in the specification of the resetting mechanism.

Let (y_n) be a wide sense stationary ARMA (p, q) process satisfying the difference equation

$$A^*y = C^*e,$$

where A^*, C^* are polynomials of the shift operator of degree p and q , respectively.

CONDITION 4.3. A^*, C^* are stable, i.e., all roots of the equations $A^*(z^{-1}) = 0, C^*(z^{-1}) = 0$ lie inside the unit circle and they are relative prime. We assume that the leading coefficients of A^* and C^* are equal to 1. The remaining coefficients of A^* and C^* are collected in a vector $\theta^* \in \mathbb{R}^{p+q}$.

To describe the noise process we assume that we are given probability space (Ω, \mathcal{F}, P) , an increasing family of σ -algebras (\mathcal{F}_n) , and a decreasing family of σ -algebras $(\mathcal{F}_n^+), n \geq 0$, such that \mathcal{F}_n and \mathcal{F}_n^+ are independent for all n , \mathcal{F}_n denotes the past before n , and \mathcal{F}_n^+ denotes the future after the moment $n + 1$, respectively.

CONDITION 4.4. The input noise process (e_n) is a wide sense stationary martingale difference process with respect to \mathcal{F}_n , i.e., $E(e_n | \mathcal{F}_{n-1}) = 0$ for $n \geq 1$. Moreover $E(e_n^2 | \mathcal{F}_{n-1}) = \sigma^2$ almost surely for all n . Finally (e_n) is L^+ -mixing with respect to $(\mathcal{F}_n, \mathcal{F}_n^+)$ and (e_n^2) is in class M^* .

Let $D \subset \mathbb{R}^{p+q}$ denote a bounded open set of system parameters θ such that the corresponding polynomials A, C , are stable and $\theta^* \in D$. For fixed $\theta \in D$ define the process $\bar{\varepsilon}(\theta)$ by the difference equation $C\bar{\varepsilon} = Ay$, i.e., $\varepsilon = (A/C)(C^*/A^*)e$. (We put $\bar{\varepsilon}_n = y_n = 0$ for $n \leq 0$). The asymptotic cost function then will be defined by

$$(4.6) \quad W(\theta) = \lim_{n \rightarrow \infty} \frac{1}{2} E \varepsilon_n^2(\theta).$$

(Note that $\bar{\varepsilon}_n(\theta)$ is not wide sense stationary due to the initial condition $\bar{\varepsilon}_n(\theta) = y_n = 0$, hence we do need a limiting procedure in (4.6).) It is well known that

$$\frac{\partial}{\partial \theta} W(\theta)|_{\theta=\theta^*} = 0 \quad \text{and} \quad \frac{\partial^2}{\partial \theta^2} W(\theta)|_{\theta=\theta^*} > 0,$$

i.e., the Hessian-matrix $W_{\theta\theta}(\theta^*)$ is positive definite.

The recursive prediction error estimator $\hat{\theta}_N$ of θ^* can be defined as follows (c.f. Söderström [37], Ljung and Söderström [33], Caines [3], or Söderström and Stoica [38]). Let $\hat{\theta}_0 \in D$ be an initial guess and set $\varepsilon_n = y_n = 0$ for $n \leq 0$. Assuming that the processes $\hat{\theta}_n, \varepsilon_n$ have been generated for $n \leq N - 1$ we define ε_N by the equation

$$(4.7) \quad (\hat{C}_{N-1}\varepsilon)_N = (\hat{A}_{N-1}y)_N.$$

Here the left-hand side of (4.7) means that the linear filter corresponding to \hat{C}_{N-1} acts on the process ε and the evaluation is done at time N . The right-hand side is interpreted similarly. $\hat{A}_{N-1}, \hat{C}_{N-1}$ denote the polynomials corresponding to $\hat{\theta}_{N-1}$. Similarly we define $(\partial/\partial\theta)\varepsilon_N$ by

$$(4.8) \quad (\hat{C}_{N-1} \frac{\partial}{\partial\theta} \varepsilon)_N = -\phi_{N-1}$$

where

$$\phi_{N-1} = (-y_{N-1}, \dots, y_{N-p}, \varepsilon_{N-1}, \dots, \varepsilon_{N-q})^T.$$

Finally let \hat{R}_{N-1} be an estimation of $(\partial^2/\partial\theta^2)W(\theta^*)$ with initial guess, say, \hat{R}_0 . Then $\hat{\theta}_N, \hat{R}_N$ are computed by the following recursion. First compute the tentative values of $\hat{\theta}_N, \hat{R}_N$ given by

$$(4.9) \quad \hat{\theta}_{N-} = \hat{\theta}_{N-1} - \frac{1}{N} \hat{R}_{N-1}^{-1} \frac{\partial}{\partial\theta} \varepsilon_N \cdot \varepsilon_N.$$

$$(4.10) \quad \hat{R}_{N-} = \hat{R}_{N-1} + \frac{1}{N} \left(\left(\frac{\partial}{\partial\theta} \varepsilon_N \right) \left(\frac{\partial}{\partial\theta} \varepsilon_N \right)^T - \hat{R}_{N-1} \right).$$

Then these tentative values will further be adjusted if they violate a certain boundedness conditions as follows. Let $D_\theta \subset D$ and D_R be compact domains in \mathbb{R}^{p+q} and $\mathbb{R}^{p \times p}$, respectively. Then we define $(\hat{\theta}_N, \hat{R}_N) = (\hat{\theta}_{N-}, \hat{R}_{N-})$ if $(\hat{\theta}_{N-}, \hat{R}_{N-}) \in D_\theta \times D_R$ and $(\hat{\theta}_N, \hat{R}_N) = (\hat{\theta}_0, \hat{R}_0)$ if $(\hat{\theta}_{N-}, \hat{R}_{N-}) \notin D_\theta \times D_R$. Note that the time is not reset!

Remark. Note that we use \hat{R}_{N-1}^{-1} in (4.9) instead of the usual term \hat{R}_N^{-1} . The form presented here makes the general result directly applicable. However, it is easy to see that the analysis given below can be carried out even if we use \hat{R}_N^{-1} in (4.9).

The domain D_θ should be chosen in such a way that the exponential stability of the time-varying filter (4.7), (4.8) is ensured. This will be achieved by imposing Condition 4.5 below. The set D_θ is a set in \mathbb{R}^{p+q} . Let the projection of D_θ on \mathbb{R}^q be denoted by D_c . For each $c \in D_c$ there corresponds a polynomial $C(z^{-1})$ and to this we can associate a companion matrix

$$\tilde{C} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ -c_1 & -c_2 & \cdots & -c_q \end{pmatrix}.$$

The set of these companion matrices will be denoted by $D_{\tilde{C}}$.

CONDITION 4.5. The matrices \tilde{C} in $D_{\tilde{C}}$ are jointly stable, i.e., there exists a symmetric positive-definite matrix U such that $\tilde{C}^T U \tilde{C} < \lambda U$ with some $0 < \lambda < 1$.

Remark. While this condition is certainly restrictive it is inherent in Ljung's scheme as we understand it now. The reason for this is that we have to assume a priori that the time-varying filters (4.7), (4.8) are exponentially stable. In an earlier attempt [25] to analyze recursive estimators, the exponential stability of (4.7) and (4.8) is taken for granted, but it is well known that the exponential stability of the frozen-parameter system, say, $C\varepsilon = Ay$, does not imply the exponential stability of the time-varying system (4.7). Whether the local analysis of Ljung's scheme given here can be "globalized" is an interesting open question.

Obviously we should also assume that $\theta^* \in D_\theta$ but this and more will be implied by Condition 4.6. As for D_R we assume that it is a compact domain of symmetric positive definite matrices such that $R^* \in \text{int} D_R$, but again more will be required by Condition 4.6.

To further specify the properties of D_θ and D_R we consider the associated ordinary differential equation

$$(4.11) \quad \dot{\theta}_t = -R_t^{-1} \frac{\partial}{\partial \theta} W(\theta_t)$$

$$(4.12) \quad \dot{R}_t = G(\theta_t) - R_t$$

where

$$G(\theta) = \lim_n \mathbf{E} \left(\frac{\partial}{\partial \theta} \bar{\varepsilon}_n(\theta) \right)^T \left(\frac{\partial}{\partial \theta} \bar{\varepsilon}_n(\theta) \right).$$

The right-hand side of this ordinary differential equation is defined in $D \times \mathbf{R}^+(p \times p)$, where $\mathbf{R}^+(p \times p)$ denotes the set of symmetric positive definite $p \times p$ matrices. It is well known that (4.11), (4.12) has a unique stationary point (θ^*, R^*) and that this equilibrium point is asymptotically stable. The last proposition is obtained by a simple eigenvalue test. It is essential for our analysis that the solution trajectories of (4.11), (4.12) starting from $(\hat{\theta}_0, \hat{R}_0)$ do not hit the boundary of $D_\theta \times D_R$. This can be ensured by the following condition.

CONDITION 4.6. Let D_R be a compact set of symmetric positive-definite matrices. We assume that $D_\theta \times D_R$ is a domain of attraction for (4.11), (4.12), i.e., for any initial value $(\theta(0), R(0)) \in D_\theta \times D_R$, the solution (θ_t, R_t) of (4.11) and (4.12) converges to (θ^*, R^*) . Furthermore we assume that $(\theta^*, R^*) \in \text{int} D_{\theta, R}$ and $(\hat{\theta}_0, \hat{R}_0) \in \text{int} D_{\theta, R}$, where $D_{\theta, R}$ is a compact domain invariant for (4.11), (4.12), and even the following stronger condition holds: the image of $D_{\theta, R}$ under the flow ϕ_t defined by (4.11), (4.12) is in $\text{int} D_{\theta, R}$ for all $t > 0$ and $D_{\theta, R} \subset D_\theta \times D_R$.

THEOREM 4.2. Under Conditions 4.3–4.6 we have

$$\hat{\theta}_N - \theta^* = O_M(N^{-1/2}) \quad \text{and} \quad (\hat{R}_N - R^*) = O_M(N^{-1/2}).$$

Proof of Theorem 4.2. To prove Theorem 4.2 we have to verify the conditions of Theorem 4.1. Conditions 4.1, 4.2 are direct consequences of the conditions of Theorem 4.2 with the following choices: $x = \theta$, $\phi = (\varepsilon, \varepsilon_\theta)$, $Q(\phi) = \varepsilon_\theta \varepsilon$, $D_0 = D_\theta \times D_R$, $D_{00} = D_{\theta, R}$. Finally, the missing part of Condition 1.5 can be derived using the fact that (4.11), (4.12) is asymptotically stable at (θ^*, R^*) and the top Lyapunov exponent can be chosen to be $-1 + c$ with any $c > 0$ since the Jacobian matrix of (4.11), (4.12) at (θ^*, R^*) has the structure

$$\begin{pmatrix} -I & 0 \\ X & -I \end{pmatrix},$$

i.e., all eigenvalues are equal to -1. (Some of the Jordan blocks may be nontrivial, though).

It follows that the stability condition (Condition 1.5) is satisfied with $\alpha = -1 + c$ with any $c > 0$ in a small invariant neighborhood of (θ^*, R^*) and thus by a compactness argument also in the whole $D_\theta \times D_R$. \square

This theorem is fundamental in the theory of identification, not only in its own right but due to its applicability to derive further very fine asymptotic results. Theorem 4.2 proved to be instrumental in proving a basic result on the closeness of recursive (on-line) and nonrecursive (off-line) estimator of ARMA-parameters expressed in the following strong approximation theorem, in which $\hat{\theta}_N$ denotes the off-line estimator of θ^* at time N .

THEOREM 4.3 (Gerencsér [13]). *Under the conditions of Theorem 4.2 we have*

$$\hat{\theta}_N - \hat{\hat{\theta}}_N = O_M(\log N/N).$$

Since the off-line estimator is much easier to analyse and is very accurately characterized in Gerencsér [17], Theorem 4.3 provides us with a substantial insight into the nature of recursive estimator processes.

Theorem 4.3 was in turn instrumental in deriving the first and so far the only real time computable criterion for model selection, i.e., order estimation of ARMA systems. Namely, in Gerencsér [14] we have shown the validity of the following theorem.

THEOREM 4.4. *Under the conditions of Theorem 4.2 we have*

$$\lim_{N \rightarrow \infty} \sum_{n=1}^N (\varepsilon_n^2 - e_n^2) / \sigma^2 \log N = (p + q), \quad \text{a.s.}$$

The important feature of this theorem is that it remains valid if one of the minimal orders p or q is overestimated. A detailed exposition of the history of this problem and the potentials of Theorem 4.4 is given in Gerencsér and Rissanen [20].

5. The proof of Theorem 4.1.

Proof of Theorem 4.1. Set

$$(5.1) \quad H(n, x, \omega) = Q(\bar{\phi}_{n+1}(x)) \quad \text{and} \quad \delta \tilde{H}(n, \omega) = Q(\phi_{n+1}) - Q(\bar{\phi}_{n+1}(x)).$$

We shall verify the conditions of Theorem 1.2. The verification of Condition 1.1 is a routine exercise using Lemma 2.4 of Gerencsér [12]. Condition 1.3 is satisfied since $EQ(\bar{\phi}_{n+1}(x)) = G(x) + O_M(\lambda^n)$ with some $0 < \lambda < 1$. The latter error term is due to “nonstationary initial conditions.” Conditions 1.4 and 1.5 are satisfied by assumption. We shall verify Condition 1.2 in Lemmas 5.1–5.3 and Condition 1.6 in Lemmas 5.4–5.5. Let us consider the process $\bar{\phi}_n = \bar{\phi}_n(x)$ generated by the filter (4.3) for fixed x .

LEMMA 5.1. *Under Conditions 4.1–4.3 the process $|\bar{\phi}_n|^2$ is in class M^* .*

Proof. Stability of the filter (4.2) implies

$$(5.2) \quad M_m(\bar{\phi}) \leq CM_m(e)$$

for all $1 \leq m \leq \infty$, where C is independent of m . Since e^2 is in class M^* , we have by Lemma 7.6

$$M_{2m}|e| \leq CM_m(|e|^2) \leq C(1 + m)$$

and thus by (5.2)

$$M_m(|\bar{\phi}|^2) \leq M_{2m}|\bar{\phi}| \leq C'(1 + m),$$

which implies the proposition. \square

Let us now define

$$\bar{\phi}_n^*(\omega) = \sup_{x \in D_0} |\bar{\phi}_n(x, \omega)|.$$

LEMMA 5.2. *The process $(\bar{\phi}_n^*)^2$ is in class M^* .*

Proof. From the proof of the previous lemma it is obvious that the process $\bar{\phi}_n(x, \omega)$ as a process parametrized by x satisfies

$$M_m(\bar{\phi}) \leq C(1 + m^{1/2}).$$

Similarly it can be shown that the process

$$\Delta \bar{\phi}_n(x, x+h, \omega) = (\bar{\phi}_n(x+h, \omega) - \bar{\phi}_n(x))/|h|, \quad h \neq 0$$

satisfies

$$M_m(\Delta \bar{\phi}) \leq C(1 + m^{1/2}).$$

Hence by the maximal inequality given in Theorem 7.2 in Appendix II we conclude that for any $1 \leq m < \infty, r > p$,

$$M_m(\bar{\phi}^*) \leq CM'_{mr}(\bar{\phi}) \leq C'(1 + m^{1/2}),$$

where C depends only on r, p , and the domains D_0 and D , and thus the proposition follows. \square

Since Q is quadratic we have

$$|Q(\bar{\phi}_n(x))| \leq C(1 + |\bar{\phi}_n(x)|)^2 \leq C(1 + \bar{\phi}_n^*)^2$$

hence $H(n, x, \omega) = Q(\bar{\phi}_{n+1}(x, \omega))$ is in class M^* .

LEMMA 5.3. *The process $\Delta H/\Delta x = (Q(\bar{\phi}_n(x+h)) - Q(\bar{\phi}_n(x)))/|h|$ is in class M^* .*

Proof. We have

$$(5.3) \quad |Q(\bar{\phi}_n(x)) - Q(\bar{\phi}_n(x'))| = \left| \int_0^1 Q_\phi(\bar{\phi}_n^\lambda)(\bar{\phi}_n(x) - \bar{\phi}_n(x')) d\lambda \right|,$$

where $Q_\phi(\phi) = (\partial/\partial\phi)Q(\phi)$ and $\bar{\phi}_n^\lambda = \lambda\bar{\phi}_n(x) + (1-\lambda)\bar{\phi}_n(x')$. Obviously $|\bar{\phi}_n^\lambda| \leq \bar{\phi}_n^*$, and since Q_ϕ is linear (5.3) is majorated by

$$(5.4) \quad C(1 + \bar{\phi}^*)|\bar{\phi}(x) - \bar{\phi}(x')|.$$

Furthermore we have

$$(5.5) \quad |\bar{\phi}_n(x) - \bar{\phi}_n(x')| = \left| \int_0^1 \bar{\phi}_{xn}(x^\lambda)(x - x') d\lambda \right|,$$

where $\bar{\phi}_{xn}(x) = (\partial/\partial x)\bar{\phi}_n(x)$ and $x^\lambda = \lambda x + (1-\lambda)x'$. Setting

$$\bar{\phi}_x^* = \sup_{x \in D_0} \|\bar{\phi}_x(x)\|$$

we get, combining (5.3), (5.4), and (5.5),

$$|Q(\bar{\phi}(x)) - Q(\bar{\phi}(x'))| \leq C(1 + \bar{\phi}^*)\bar{\phi}_x^*|x - x'|.$$

$(\bar{\phi}_{x,n}^*)^2$ can be shown to be in class M^* . To complete the proof we use Lemma 7.7. \square

The most difficult task is the verification of Condition 1.6, which we now begin. Note that we have

$$\delta\tilde{H}(n, \omega) = Q(\phi_{n+1}) - Q(\bar{\phi}_{n+1}(x_n)).$$

LEMMA 5.4. *We have*

$$|Q(\phi_{n+1}) - Q(\bar{\phi}_{n+1}(x_n))| \leq Q_{\phi_n}^* |\phi_{n+1} - \bar{\phi}_{n+1}(x_n)|,$$

where $Q_{\phi_n}^*$ is an M -bounded sequence.

Proof. We have

$$|Q(\phi_{n+1}) - Q(\bar{\phi}_{n+1}(x_n))| \leq \int_0^1 Q_\phi(\phi_{n+1}^\lambda)(\phi_{n+1} - \bar{\phi}_{n+1}(x_n)) d\lambda,$$

where $\phi_{n+1}^\lambda = \lambda\phi_{n+1} + (1-\lambda)\bar{\phi}_{n+1}(x_n)$. Here $Q_\phi(\phi_{n+1}^\lambda)$ is M -bounded. To see this it is sufficient to show that ϕ_{n+1} and $\bar{\phi}_{n+1}(x_n)$ are M -bounded. Now ϕ_{n+1} is M -bounded by Conditions 4.1 and 4.2, while $|\bar{\phi}_{n+1}(x_n)|$ is majorated by $\bar{\phi}_{n+1}^*$. Thus the lemma is proved. \square

LEMMA 5.5. *We have*

$$|\phi_{n+1} - \bar{\phi}_{n+1}(x_n)| \leq \frac{1}{n}\rho_n + \sum_{i=1}^n K\lambda^{n-i}\chi_{B_i}$$

where (ρ_n) is M -bounded, K is a positive constant, and $0 < \lambda < 1$.

Proof. We have

$$(5.6) \quad \phi_{n+1} - \bar{\phi}_{n+1}(x_n) = A(x_n)(\phi_n - \bar{\phi}_n(x_{n-1})) + A(x_n)(\bar{\phi}_n(x_{n-1}) - \bar{\phi}_n(x_n))$$

with an M -bounded initial condition. Using Condition 4.1 and the inequality

$$|\phi_n(x) - \phi_n(x')| \leq C\bar{\phi}_{x_n}^* \cdot |x - x'|,$$

we get

$$(5.7) \quad |\phi_{n+1} - \bar{\phi}_{n+1}(x_n)| \leq O_M(\lambda^{n+1}) + C \sum_{i=1}^n \lambda^{n-i} \bar{\phi}_{x_i}^* |x_i - x_{i-1}|.$$

On the other hand we have

$$(5.8) \quad |x_{i+1} - x_i| \leq \frac{1}{i+1} Q(\phi_{i+1}) + K\chi_{B_{i+1}},$$

where K is an upper bound for the diameter of the set D . Since $Q(\phi_{n+1})$ is M -bounded the effect of the first term of (5.8) onto $\phi_{n+1} - \bar{\phi}_{n+1}(x_n)$ is $O_M(n^{-1})$ (c.f. Lemma 7.5 of Appendix II). Substituting (5.8) into (5.7) gives the lemma. \square

LEMMA 5.6. *We have with some $\varepsilon > 0$*

$$\sup_{s \leq \sigma \leq qs} \sum_{r=\sigma}^{t(\sigma) \wedge qs} \frac{1}{r} |\delta\tilde{H}(r, \omega)| = O_M(s^{1/2-\varepsilon}).$$

Proof. Combining the last two lemmas we get for

$$\delta H(n, \omega) = Q(\phi_{n+1}) - Q(\bar{\phi}_{n+1}(x_n))$$

the inequality

$$|\delta H(n, \omega)| \leq Q_{xn}^* \left(\frac{1}{n} \rho_n + \sum_{i=1}^n K \lambda^{n-i} \chi_{B_i} \right).$$

Let us now consider a subdivision of the sequence of integers by the points $s = 2^i$, and for any $s \geq 2$ consider the “intervals” $[s/2, 2s]$.

We have

$$\sum_{r=\sigma}^{\tau(\sigma) \wedge qs} \frac{1}{r} |\delta H(r, \omega)| \leq \sum_{r=\sigma}^{\tau(\sigma) \wedge qs} \frac{1}{r} Q_{xr}^* \left(\frac{1}{r} \rho_r + \sum_{i=1}^r K \lambda^{r-i} \chi_{B_i} \right).$$

The contribution of the first term in the bracket is majorated by

$$\sum_{r=s}^{qs} \frac{1}{r} Q_{xr}^* \frac{1}{r} \rho_r = O_M(s^{-1}).$$

The contribution of the second term in the bracket can be written after changing the order of summation as

$$\sum_{i=1}^{\sigma} C \chi_{B_i} \sum_{r=\sigma}^{\tau(\sigma) \wedge qs} \lambda^{r-i} \frac{1}{r} Q_{xr}^*.$$

We can majorate the last sum by

$$\frac{1}{s} \left(\sum_{i=1}^{\sigma} C \chi_{B_i} \lambda^{\sigma-i} \right) \left(\sum_{r=\sigma}^{\tau(\sigma) \wedge qs} \lambda^{r-\sigma} Q_{xr}^* \right) \leq \frac{K}{s(1-\lambda)} \sum_{r=\sigma}^{\tau(\sigma) \wedge qs} \lambda^{r-\sigma} Q_{xr}^*.$$

Now Q_{xr}^* is an L^+ -mixing process since the class of L^+ -mixing processes is invariant under the operations we use to get Q_{xr}^* from the input noise e_n (c.f. Lemmas 7.8–7.10 of Appendix II). Hence by Theorem 6.1 we have for some $\varepsilon > 0$

$$Q_{x,qs}^{**} \triangleq \sup_{1 \leq \sigma < qs} \sum_{r=\sigma}^{\tau(\sigma) \wedge qs} \lambda^{r-\sigma} Q_{xr}^* = O_M((qs)^{1/2-\varepsilon}).$$

Thus we get the lemma. \square

6. Appendix I. Estimation of extreme values. The objective of this section is to derive the result we used in completing the proof of Lemma 5.6.

THEOREM 6.1. *Let (u_n) be a nonnegative L^+ -mixing process and define the process*

$$X_n = \sum_{r=n}^{\infty} \lambda^{r-n} u_r$$

with some $0 < \lambda < 1$. Then we have for some $\varepsilon > 0$

$$\sup_{0 \leq n \leq N} X_n = O_M(N^{1/2-\varepsilon}).$$

To prepare the proof we first consider a simpler problem, the estimation of the supremum of independent random variables.

THEOREM 6.2. *Let (X_n) be a sequence of independent M -bounded random variables, and let*

$$Z_N = \max_{n \leq N} |X_n|.$$

Then for any $\varepsilon > 0$

$$EZ_N \leq 3M_{1/\varepsilon}(X)N^\varepsilon.$$

Remark. Although an extensive literature is available on extremal processes (c.f. Galambos [10]), the above theorem seems to be new.

Proof. Let the distribution function of $|X_n|$ be $F_n(x)$. Then

$$P(Z_N \leq x) = \prod_{n=1}^N F_n(x).$$

Since $(|X_n|)$ is M -bounded, we have for any $m \geq 1$ and $x > 0$,

$$1 - F_n(x) \leq E|X_n|^m / x^m,$$

or equivalently for any $m \geq 1$ and $x > 0$ we have $F_n(x) \geq 1 - E|X_n|^m / x^m$, hence if the last lower bound is positive we get

$$(6.1) \quad P(Z_N \leq x) \geq (1 - M_m^m(X)x^{-m})^N.$$

We need a lower bound for $P(Z_N \leq x)$, hence we estimate the function $(1 - z)^N$ from below. Using the inequality $\log(1 - z) \geq -z - z^2 > -2z$ for $0 < z < 1/2$, we get $N \log(1 - z) \geq -2Nz$.

Now if $2Nz < 1$, then we have $-2Nz > \log(1 - 2Nz)$. Hence we finally get: for $2Nz < 1$ we have

$$(1 - z)^N \geq 1 - 2Nz,$$

or equivalently

$$(6.2) \quad 1 - (1 - z)^N \leq 2Nz.$$

Now we can estimate the tail probabilities $P(Z_N > x)$ as follows: from (6.1) we have

$$(6.3) \quad P(Z_n > x) \leq 1 - (1 - M_m^m(X)x^{-m})^N,$$

hence (6.2) gives, with $z = M_m^m(X)x^{-m}$, the estimate

$$(6.4) \quad P(Z_n > x) \leq 2NM_m^m(X)x^{-m},$$

whenever $2Nz < 1$, i.e., whenever $x > 2^{1/m}N^{1/m}M_m(X)$.

Let us now use the identity

$$EZ_N = \int_0^\infty P(Z_N > x)dx.$$

Let $x_1 = 2M_m(X)N^{1/m}$. Then we have

$$(6.5) \quad \int_0^{x_1} P(Z_N > x) dx \leq x_1 = 2M_m(X)N^{1/m}.$$

On the other hand in the region $x > x_1$ we have

$$\int_{x_1}^{\infty} P(Z_N > x) dx \leq \int_{x_1}^{\infty} 2NM_m^m(X)x^{-m} dx,$$

and here the right-hand side equals $NM_m^m(X)(1/(m-1))x_1^{-m+1}$. Substituting $x_1 = 2M_m(X)N^{1/m}$ we get after trivial arithmetic

$$(6.6) \quad \int_{x_1}^{\infty} (1 - P(Z_N \leq x)) dx \leq M_m(X)N^{1/m}.$$

Adding (6.5) and (6.6) we get the proposition of the theorem. \square

Proof of Theorem 6.1. Let us set $k = [N^\varepsilon]$ and approximate X_n by

$$X_{0n} = \sum_{r=n}^{n+k-1} \lambda^{r-n} u_r.$$

Then $|X_n - X_{0n}| = O_M(\lambda^k)$, hence

$$\left| \sup_{n \leq N} X_n - \sup_{n \leq N} X_{0n} \right| = O_M(N\lambda^k) = O_M(1).$$

Now for $l = 1, \dots, k$ define the sets $I_l = \{n : n = l + km, n \leq N\}$ with m integer. For each $n = l + km$ we approximate X_{0n} by $X_{0n}^+ = E(X_{0n} | \mathcal{F}_{l+k(m-1)}^+)$. We have

$$X_{0n} - X_{0n}^+ = \sum_{r=n}^{n+k-1} \lambda^{r-n} u_{n,l+k(m-1)}^+$$

where $u_{n,n'}^+ = E(u_n | \mathcal{F}_{n'}^+)$ for $0 \leq n' \leq n$. Taking the $L_q(\Omega, \mathcal{F}, P)$ -norm on both sides and using the quasi monotonicity of $\gamma_q(\tau, u)$, which means $\gamma_q(\tau', u) < 2\gamma_q(\tau, u)$ for $\tau' > \tau$, we get

$$E^{1/q} |X_{0n} - X_{0n}^+|^q \leq \sum_{r=n}^{n+k-1} 2\lambda^{r-n} \gamma_q(k, u) \leq 2/(1-\lambda) \gamma_q(k, u).$$

Therefore we get

(6.7)

$$E^{1/q} \left| \sup_{n \leq N} X_{0n} - \sup_{n \leq N} X_{0n}^+ \right|^q \leq N X_{0n}^+{}^q \leq E^{1/q} \left(\sum_{r \leq N} |X_{0n} - X_{0n}^+| \right)^q \leq 2N/(1-\lambda) \cdot C_q/k^{1+c}.$$

On the other hand, the sequence $X_{0n}^+, n \in I_l$, is independent, and $E^{1/q} |X_{0n}^+|^q \leq 2M_q(X_0) < \infty$ for all n and l . Therefore for any $q \geq 1$ and $\delta > 0$,

$$E^{1/q} \left| \sup_{n \in I_l} X_{0n}^+ \right|^q \leq 6M_{q/\delta}(X_0)(N/k)^\delta,$$

since $|I_l| \leq N/k$. Finally we get

$$E^{1/q} \left| \sup_{n \leq N} X_{0n}^+ \right|^q \leq \sum_{l=1}^k E^{1/q} \left| \sup_{n \in I_l} X_{0n}^+ \right|^q \leq 6M_{q/\delta}(X_0)k(N/k)^\delta.$$

Combining this inequality with (6.7) we get

$$(6.8) \quad E^{1/q} \left| \sup_{n \leq N} X_{0n} \right|^q \leq \frac{2N}{1-\lambda} \frac{C_q}{k^{1+c}} + 6M_{q/\delta}(X_0)k(N/k)^\delta.$$

To find the optimal choice of k consider the function $ax^\alpha + bx^{-\beta}$, $x > 0$, with $a, b, \alpha, \beta > 0$. It is easy to see that this function is minimized at $x = (b\beta/a\alpha)^{1/(\alpha+\beta)}$ and the minimal value is

$$a^{\beta/(\alpha+\beta)} b^{\alpha/(\alpha+\beta)} \left(\left(\frac{\beta}{\alpha} \right)^{\alpha/(\alpha+\beta)} + \left(\frac{\alpha}{\beta} \right)^{\beta/(\alpha+\beta)} \right).$$

Choosing $k = x$, $\alpha = 1 - \delta$, $\beta = 1 + c$, $a = CN^\delta$, and $b = CN$, where C is a system constant, we get that the optimal k is given by $k = CN^{(1-\delta)/2+c-\delta}$, and the minimal value of the right-hand side of (6.8) is

$$C(N^{\delta(1+c)}N^{1-\delta})^{1/(2+c-\delta)} = CN^{(1+c)/(2+c-\delta)} = CN^{1/2-\varepsilon}$$

with some $\varepsilon > 0$ for sufficiently small δ . Thus the theorem has been proved. \square

7. Appendix II. Auxiliary results. First we present two results published in (Gerencsér [12]) and used in this paper. The following moment inequality follows from Theorems 1.1 and 4.1 of the cited paper.

THEOREM 7.1. *Let (u_t) , $t \geq 0$ be an L -mixing process with $Eu_t = 0$ for all t . Let (f_t) be a function in $L_2[0, T]$. Define*

$$I_{a,b}^*(f) = \sup_{a \leq t \leq b} \left| \int_a^t f_s u_s ds \right|.$$

Then we have for all $m > 2$

$$E^{1/m} |I_{a,b}^*(f)|^m \leq C'_m \left(\int_a^b f_t^2 dt \right)^{1/2} M_{2m}^{1/2}(u) \cdot \Gamma_{2m}^{1/2}(u),$$

where C_m is independent of a, b .

Let us consider a stochastic process $(u_n(\theta))$, which is measurable, separable, M -bounded, and M -Lipschitz continuous in θ , with exponent α for $\theta \in D$. By Kolmogorov's theorem the realizations of $(x_n(\theta))$ are continuous in θ with probability 1, hence we can define for almost all ω

$$u_n^* = \max_{\theta \in D_0} |u_n(\theta)|,$$

where $D_0 \subset \text{int} D$ is a compact domain.

THEOREM 7.2 (Theorem 3.4 in Gerencsér [12]). *Assume that $(u_n(\theta))$ is a stochastic process which is measurable, separable, M -bounded, and M -Lipschitz continuous*

in θ for $\theta \in D$. Let u_n^* be the random variable defined above. Then we have for all positive integers q and $s > p$,

$$M_q(u^*) \leq C(M_{qs}(u) + M_{qs}(\Delta u / \Delta^\alpha \theta)),$$

where C depends only on p, q, s , and D_0, D .

LEMMA 7.3. Let $(v_t(y))$, $y \in D$, be an L -mixing process, uniformly in y with respect to a family of σ -algebras $(\mathcal{F}_t, \mathcal{F}_t^+)$, $t \geq 0$. Here $D \subset \mathbb{R}^p$ is an open domain. Assume that (y_t) , $t \geq 0$, is a measurable function taking values in D . Then the process $u_t = v_t(y_t)$ is L -mixing with respect to $(\mathcal{F}_t, \mathcal{F}_t^+)$ and we have for any $m \geq 1$

$$M_m(u) \leq M_m(v) \quad \Gamma_m(u) \leq \Gamma_m(v).$$

Proof. Let $t > s$ and let us approximate $v_t(y_t)$ by $v_{t,s}^+(y_t)$, where $v_{t,s}^+(y) = E(v_t(y) | \mathcal{F}_s^+)$. Obviously $v_{t,s}^+(y_t)$ is \mathcal{F}_s^+ -measurable since y_t is deterministic. Furthermore we have for any $q \geq 1$

$$E^{1/q} |v_t(y_t) - v_{t,s}^+(y_t)|^q \leq \gamma_q(t - s, y),$$

which implies the claim of the lemma. \square

It is easy to see that exponentially stable linear filters are input-output M -bounded in the sense that the response to an M -bounded input is an M -bounded output. We present two lemmas below showing that certain rates of growth of the input process are also preserved by an exponentially stable linear filter.

LEMMA 7.4. Let (u_n) , $n \geq 0$, be an M -bounded process and define a process (x_n) by

$$(7.1) \quad x_{n+1} = \lambda x_n + \rho^n u_n, \quad x_0 = 0,$$

where $0 < \lambda < \rho < 1$. Then for any $m \geq 1$ we have

$$E^{1/m} |x_n|^m \leq \frac{\rho^n}{\rho - \lambda} M_m(u).$$

On the other hand, if $0 < \rho < \lambda < 1$, then we have

$$E^{1/m} |x_n|^m \leq \frac{\lambda^{n-1}}{(1 - \lambda)} \cdot M_m(u).$$

Proof. Let $z_n = \rho^{-n} x_n$. Then we have, after multiplying (7.1) by $\rho^{-(n+1)}$,

$$z_{n+1} = \lambda \rho^{-1} z_n + \rho^{-1} u_n,$$

which can be solved explicitly for z_n :

$$z_n = \sum_{i=0}^{n-1} (\lambda \rho^{-1})^{n-1-i} \rho^{-1} u_i.$$

Using the triangle inequality for the $L_m(\Omega, \mathcal{F}, P)$ -norm and the condition $0 < \lambda < \rho$, we get

$$M_m(z) \leq (1 - \lambda \rho^{-1})^{-1} \rho^{-1} M_m(u),$$

from which the proposition follows.

In the case when $0 < \rho < \lambda$, we define $z_n = \lambda^{-n}x_n$, and get

$$z_{n+1} = \lambda z_n + \lambda^{-n-1}\rho^n u_n.$$

From here

$$z_n = \sum_{i=0}^{n-1} \lambda^{n-1-i}(\lambda^{-i-1}\rho^i u_i).$$

Taking the $L_m(\Omega, \mathcal{F}, P)$ -norm of both sides, taking into account that $(\lambda^{-i-1}\rho^i) \leq \lambda^{-1}$ and using the triangle inequality we get the claim of the lemma. \square

LEMMA 7.5. Let $v_i, i = 1, 2, \dots$, be an \mathbb{R} -valued stocastic process such that $v_i = O_{M(g_i)}$, where $g_i > 0$ satisfies $\lim_{i \rightarrow \infty} g_i/g_{i-1} = 1$. Define x by

$$x_N = \lambda x_{N-1} + v_i, \quad x_0 = 0,$$

where $|\lambda| < 1$. Then $x_N = O_M(g_N)$.

Proof. Set $z_N = g_N^{-1}x_N$, then z_N satisfies

$$(7.2) \quad z_N = \lambda g_{N-1}g_N^{-1}z_{N-1} + g_N^{-1}v_{N-1}, \quad z_0 = 0.$$

Since $\overline{\lim}|\lambda g_{N-1}g_N^{-1}| < 1$ and $g_N^{-1}v_{N-1} = O(1)$, the proposition follows by solving (7.2) for z_N and applying the triangle inequality. \square

Now we present some simple estimates related to the class M^* .

LEMMA 7.6. Let $\kappa = \kappa(\omega)$ be a random variable. Then $E \exp \varepsilon \kappa < \infty$ for some $\varepsilon > 0$ if and only if we have, with some $C > 0$, and all $m \geq 1$.

$$(7.3) \quad M_m(\kappa) \leq C(1 + m).$$

Proof. Note that $E \exp \varepsilon \kappa$ is finite if and only if $E \exp \varepsilon |\kappa|$ is finite. Furthermore

$$E \exp \varepsilon |\kappa| = \sum_{m=0}^{\infty} E|\varepsilon \kappa|^m / m!$$

by the Beppo-Levi theorem. Thus if the left-hand side is finite, then with some C

$$E|\varepsilon \kappa|^m \leq C m! < C(m/e)^m m^{1/2}$$

by Stirling's formula, and (7.3) follows by taking m th root.

Conversely, if (7.3) holds then

$$E|\varepsilon \kappa|^m / m! \leq (\varepsilon C)^m (1 + m)^m / m! < C' \gamma^m$$

with $C' > 0, 0 < \gamma < 1$ if ε is sufficiently small, hence summation over m and repeated use of the Beppo-Levi theorem gives $E \exp \varepsilon |\kappa| < \infty$. \square

Remark. Similarly it can be shown that $E \exp \varepsilon \kappa^2 < \infty$ for some $\varepsilon > 0$ if and only if

$$M_m(\kappa) \leq C(1 + m^{1/2}).$$

Indeed $E \exp \varepsilon \kappa^2 < \infty$ for some $\varepsilon > 0$ if and only if

$$M_m(|\kappa|^2) = M_{2m}^2(\kappa) < C(1 + m).$$

LEMMA 7.7. *If ξ, η are random variables such that $E\xi^2 < \infty$ and $E\eta^2 < \infty$ for some $\varepsilon > 0$, then for small positive ε 's,*

$$(7.4) \quad E \exp \varepsilon \xi \eta < \infty.$$

Proof. The inequalities

$$M_m(\xi) \leq C(1 + m^{1/2}), \quad M_m(\eta) \leq C(1 + m^{1/2})$$

imply by the Cauchy-Schwartz inequality

$$M_m(\xi\eta) \leq M_{2m}(\eta) \leq C'(1 + m),$$

which implies (7.4). \square

Finally we present a few auxiliary results on L -mixing processes. The results below show that the class of L^+ -mixing processes is invariant under the operations that are used in the identification of linear stochastic systems.

LEMMA 7.8. *Let (u_n) , $n \geq 0$, be a real-valued L^+ -mixing process and define x_n by the equation*

$$x_{n+1} = \lambda x_n + u_n, \quad x_0 = 0,$$

where $|\lambda| < 1$. Then (x_n) also is L^+ -mixing.

Proof. Following the proof of Lemma 2.4 in Gerencsér [12] (c.f. especially (2.6)) we get

$$\gamma_q(\tau, x) \leq 2M_q(u)\lambda^\tau(1 - \lambda)^{-1} + 2C'(\lambda^\tau * \tau^{-1-c}) = O(\tau^{-1-c}). \quad \square$$

(Here we used Lemma 7.5 to get the final estimate).

Remark. Obviously the lemma remains valid if (u_n) is vector-valued and (x_n) is defined as the output of a finite-dimensional stable linear system, the input of which is (u_n) .

LEMMA 7.9. *Assume that (u_n) , (v_n) are L^+ -mixing processes. Then $z_n = u_n \cdot v_n$ is also L^+ -mixing.*

Proof. We have for any $\tau \geq 1$ integer and $q \geq 1$

$$\gamma_q(\tau, z) \leq \gamma_{2q}(\tau, u)M_{2q}(v) + M_{2q}(u)\gamma_{2q}(\tau, v),$$

from which the claim immediately follows. \square

LEMMA 7.10. *Let $u = (u_n(x, \omega))$ be a separable stochastic defined for $x \in D \subset \mathbb{R}^p$, where D is an open domain. Assume that u and $\Delta u / \Delta x = (u_n(x + h, \omega) - u_n(x) / |h|)$ $h \neq 0$ are L^+ -mixing uniformly in x and $(x, x + h)$, respectively. Let $D_0 \subset D$ be a compact domain and define*

$$u_n^* = \sup_{x \in D_0} |u(n, x, \omega)|.$$

Then $u^ = (u_n^*)$ is L^+ -mixing.*

Proof. We have by Theorem 7.2 for any $\tau \geq 1$ integer, $q \geq 1$ and $r > p$,

$$\gamma_q(\tau, u^*) \leq C(\gamma_{qr}(\tau, u) + \gamma_{qr}(\tau, \Delta u / \Delta x)),$$

where C depends only on p, q, r, D_0 , and D . Hence the claim trivially follows for u^* . \square

Acknowledgments. The author thanks Karim Nassiri-Taussi and Jimmy Baikovicius for their careful reading of the manuscript, and Mindle Levitt, Jimmy Baikovicius, and Solomon Seifu for their considerable amount of work in the preparation of this document. Also, the author thanks the reviewer for pointing out the possibility of significantly simplifying the proof of the main theorem (Theorem 1.1).

REFERENCES

- [1] A. BENVENISTE, M. METIVIER, AND P. PRIOURET, *Algorithms adaptatifs et approximations stochastiques*, Masson, Paris, 1987.
- [2] A. N. BORODIN, *A stochastic approximation procedure in the case of weakly dependent observations*, Theory Probab. Appl, 24 (1979), pp. 34–52.
- [3] P. E. CAINES, *Linear Stochastic Systems*, John Wiley, New York, 1988.
- [4] H. F. CHEN, *Recursive Estimation and Control for Stochastic Systems*, John Wiley, New York, 1985.
- [5] H. F. CHEN, L. GUO, AND Y. F. ZHANG, *Identification and adaptive control for ARMAX systems*, in Topics in Stochastic Systems: Modelling, Estimation and Adaptive Control, L. Gerencsér and P. E. Caines, eds., Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin, New York, 1991, pp. 216–241.
- [6] M. H. A. DAVIS AND R. B. VINTER, *Stochastic Modelling and Control*, Chapman and Hall, New York, 1985.
- [7] D. P. DJERJEVECKII AND A. L. FRADKO, *Applied Theory of Discrete Adaptive Control Systems*, Nauka, Moscow, 1981 (in Russian).
- [8] T. E. DUNCAN AND B. PASIK-DUNCAN, *Adaptive Control of Continuous Time Linear Stochastic Systems*, Mathematics of Control, Signals and Systems, 3 (1990), pp. 45–60.
- [9] ———, *Some methods for the adaptive control of continuous time linear stochastic systems*, in Topics in Stochastic Systems: Modelling, Estimation and Adaptive Control, L. Gerencsér and P. E. Caines, eds., Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin, New York, 1991, pp. 242–267.
- [10] J. GALAMBOS, *The Asymptotic Theory of Extreme Order Statistics*, John Wiley, New York, 1978.
- [11] S. GEMAN, *Some averaging and stability results for random differential equations*, SIAM J. Appl. Math., 36 (1979), pp. 87–105.
- [12] L. GERENCSÉR, *On a class of mixing processes*, Stochastics, 26 (1989), pp. 165–191.
- [13] ———, *Strong approximation of the recursive maximum likelihood estimator of the parameters of an ARMA process*, McGill Research Center for Intelligent Machines TR-CIM-89-8 (1989). Systems and Control Lett., submitted.
- [14] ———, *On Rissanen's predictive complexity for stationary ARMA processes*, McGill Research Center for Intelligent Machines TR-CIM-89-5, (1989). J. of Statistical Planning and Inference, submitted.
- [15] ———, *Closed loop parameter identifiability and adaptive control of a linear stochastic system*, Systems Control Lett., (1990), pp. 411–416.
- [16] ———, *Fixed gain stochastic approximation processes*, J. Math. Systems, Estimation and Control, (1990), submitted, under revision.
- [17] ———, *On the martingale approximation of the estimation error of ARMA parameters*, Systems Control Lett., (1990), pp. 417–423.
- [18] ———, *Strong approximation results in estimation and adaptive control*, in Topics in Stochastic Systems: Modelling, Estimation and Adaptive Control, L. Gerencsér and P. E. Caines, eds., Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin, New York, 1991, pp. 268–299.
- [19] L. GERENCSÉR, I. GYÖNGY, AND G. MICHALETZKY, *Continuous-time Recursive Maximum-Likelihood Method. A New Approach to Ljung's Scheme*, Proc. of the 9th IFAC World Congress, Budapest, Vol. 2, L. Ljung and K.J. Åström, eds., Pergamon Press, Oxford, 1984, pp. 75–77.
- [20] L. GERENCSÉR AND J. RISSANEN, *Asymptotics of predictive stochastic complexity*, in New Directions in Time Series Analysis, Proc. of the 1990 IMA Workshop, E. Parzen, D. Brillinger, M. Rosenblatt, M. Taqqu, J. Geweke, and P. E. Caines, eds., Springer-Verlag, Berlin, New York, 1992, to appear.

- [21] G. C. GOODWIN, P. Y. RAMADGE, AND P. E. CAINES, *Discrete time multi-variable adaptive control*, IEEE Trans. Automat. Control, AC-25 (1981), pp. 449–456.
- [22] G. C. GOODWIN AND K. S. SIN, *Adaptive Filtering, Prediction and Control*, Prentice-Hall, Englewood Cliffs, NJ, 1984.
- [23] P. HALL AND C. C. HEYDE, *Martingale Limit Theory and Its Applications*, Academic Press, New York, 1980.
- [24] E. J. HANNAN, *The convergence of some time-series recursions*, Ann. Statist., 4, (1976), pp. 1258–1270.
- [25] ———, *Recursive estimation based on ARMA models*, Ann. Statist., 8 (1980), pp. 1258–1270.
- [26] P. HARTMAN, *Ordinary Differential Equations*, John Wiley, New York, 1964.
- [27] A. HEUNIS, *Rates of convergence for an adaptive filtering algorithm driven by stationary dependent data*, SIAM J. Control Optim., (1991), submitted.
- [28] H. J. KUSHNER AND D. S. CLARK, *Stochastic Approximation Methods for Constrained and Unconstrained Optimization*, Springer-Verlag, Berlin, New York, 1978.
- [29] H. J. KUSHNER, *Approximation and Weak Convergence Methods for Random Processes*, MIT Press, Cambridge, MA, 1984.
- [30] T. Z. LAI, *Asymptotically efficient recursive estimation and adaptive control in stochastic regression models and ARMAX systems*, in Topics in Stochastic Systems: Modelling, Estimation and Adaptive Control, L. Gerencsér and P. E. Caines, eds., Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin, New York, 1991, pp. 335–368.
- [31] T. Z. LAI AND C. Z. WEI, *Least squares estimates in stochastic regression models with application to identification and control of dynamic systems*, Ann. Statist., 10 (1982), pp. 154–165.
- [32] L. LJUNG, *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 551–575.
- [33] L. LJUNG AND T. SÖDERSTRÖM, *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, MA, 1984.
- [34] J. B. MOORE, *On strong consistency of least squares identification algorithms*, Automatica, 14 (1978), pp. 505–509.
- [35] L. S. PONTRYAGIN, *Ordinary Differential Equations*, Nauka, Moscow, 1970 (in Russia).
- [36] J. RISSANEN, *Stochastic complexity and modeling*, Ann. Statist., 14 (1986), pp. 1080–1100.
- [37] T. SÖDERSTRÖM, *An on-line algorithm for approximate maximum-likelihood identification of linear dynamic systems*, Report 7308, Department of Automatic Control, Lund Institute of Technology, Lund, Sweden, 1973.
- [38] T. SÖDERSTRÖM AND P. STOICA, *System Identification*, Prentice-Hall, New York, 1989.
- [39] V. SOLO, *On the convergence of AML*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 958–962.
- [40] ———, *The second order properties of a time series recursion*, Ann. Statist. 9 (1981), pp. 307–317.
- [41] D. M. WIBERG, *The MIMO Wiberg Estimator*, Proc 28th IEEE Conference on Decision and Control, Vol. 3, pp. 2590–2594.
- [42] G. YIN, *Recent progress in parallel stochastic approximations*, in Topics in Stochastic Systems: Modelling, Estimation and Adaptive Control, L. Gerencsér and P. E. Caines, eds., Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin, New York, 1991, pp. 159–184.
- [43] K. YOSHIHARA, *Moment inequalities for mixing sequences*, Kodai Math. J., 1 (1978), pp. 316–328.

SUBOPTIMIZATION OF SINGULARLY PERTURBED CONTROL SYSTEMS*

VLADIMIR GAITSGORY†

Abstract. A technique different from the boundary layer method is developed to deal with singularly perturbed optimal control problems. The technique is applicable, in particular, in the case when the optimal control takes the form of “fast” oscillations and the boundary layer method cannot be used. Necessary and sufficient conditions for the optimal control to be “slow” are given. The results are illustrated by examples.

Key words. singular perturbations, optimal control, suboptimization, asymptotic properties

AMS(MOS) subject classifications. 49B10, 49B50

1. Introduction. Problems of optimal control of singularly perturbed systems have been considered by many authors (see the overviews in Bensoussan [3], Kokotovic [18], Kokotovic, Khalil, and O'Reilly [19], Kokotovic, O'Malley, and Sannuti [20], and Saksena, O'Reilly, and Kokotovic [25]). Many have been concerned, in particular, with the following matter. Given a singularly perturbed control system

$$(1.1) \quad \dot{z} = f_1(z, y, u), z(0) = z_0,$$

$$(1.2) \quad \epsilon \dot{y} = f_2(z, y, u), y(0) = y_0$$

and some cost function, it is asked whether the solution to such an optimal control problem can be approximated in some way by the solution to the problem of optimization of the so-called reduced system

$$(1.3) \quad \dot{z} = f(z, \psi(z, u), u), z(0) = z_0,$$

which is formally obtained from (1.1), (1.2) by (i) equating ϵ to zero and thus transforming (1.2) into the static equation

$$(1.4) \quad 0 = f_2(z, y, u),$$

and (ii) substituting the root $y = \psi(z, u)$ of this equation into (1.1). Here ϵ is a small positive parameter; f_1, f_2 are vector functions with their values in R^{n_1} and R^{n_2} , respectively; and $u(t)$ is a control chosen usually among measurable functions, satisfying the inclusion

$$(1.5) \quad u(t) \in U,$$

where U is a closed subset of R^m .

A heuristic explanation of the possibility that the resolution of the above matter is positive is connected with the hypothesis that the optimal control is in some sense slow and that the variables y (called fast) converge rapidly to their quasi-stationary state defined by the root of (1.4) and remain in a neighborhood of this root, while the variables z (called slow) are changing in accordance with (1.3).

* Received by the editors August 6, 1990; accepted for publication (in revised form) July 8, 1991.

† Department of Economics and Business Administration, Bar-Ilan University, 52900 Ramat-Gan, Israel.

The validity of a similar description was established in Tichonov [27] for uncontrolled singularly perturbed differential equations. In optimal control problems, the hypothesis may not be true, however. The controls and fast variables may, for instance, oscillate rapidly in the optimal regime (see Example 4.2 in §4). Equation (1.4) does not, then, describe any connection between y and z , u , and thus the control optimizing (1.3) may be far from the optimal one in (1.1), (1.2).

A traditional way of dealing with the problems of optimal control of systems (1.1), (1.2) is an application of the boundary layer method (O'Malley [21], Vasil'eva, and Butuzov [28]) to the differential systems constructed on the base of necessary or sufficient optimality conditions, with the validity of the hypothesis about "slowness" of the optimal control being guaranteed by the properties of the problems considered.

Although it allows us to find suboptimal solutions in many important cases (see the mentioned overviews), this approach does not, however, allow us to verify the applicability of the control slowness hypothesis in a general case and to approximate the solution if this hypothesis is not true. In this paper, we develop another approach (Gaitsgory [11], [12], Plotnikov [24]) which can be considered an extension of the averaging method described for uncontrolled motion in Volosov [29]. The approach is based on the hypothesis formalized as Assumption 2.1 in §2 and imposes no restrictions on the rates of changing of controls. The main idea of the approach is an approximation of z -components of the trajectories of system (1.1), (1.2) by the solutions to some differential inclusion with a convex-valued right-hand side. Also, it allows for the use of the latter to obtain the complete answer to the question about whether the optimal regime in (1.3) approximates the optimal regime in (1.1), (1.2). This answer is given in the paper in the form of necessary and sufficient conditions, which, although are difficult for a direct verification, permit us to construct new criteria for the approximation, via the optimization of (1.3), to be true and to also separate the situations when it is not the case.

The paper consists of five sections. The first is the Introduction. General statements about the approximation of system (1.1), (1.2) by the differential inclusion and the connections with the reduced system (1.3) are established under Assumption 2.1 in §2. Results that allow us to apply these statements in dealing with a sufficiently wide class of singularly perturbed optimal control problems are obtained in §3. This class of control problems is considered in §4. The most tedious proofs are gathered in §5.

2. Definitions and basic lemmas. Along with system (1.1), (1.2), consider a so-called associated system

$$(2.1) \quad \dot{y} = f_2(z, y, u), z = \text{constant},$$

which differs from system (1.2) by the replacement of the timescale $\tau = t\epsilon^{-1}$ and by the fact that the vector z is fixed at some constant level. Let us denote by $y_z(\tau, u(\cdot), y)$ the solution to system (2.1) obtained with some admissible control $u(t)$ and with initial values y . Suppose that the integral $\int_0^S f_1(z, y_z(\tau, u(\cdot), y), u(\tau)) d\tau$ exists. Divide it by S and denote by $V(z, S, y)$ the union of such integrals over all admissible controls

$$(2.2) \quad V(z, S, y) = \bigcup \left\{ S^{-1} \int_0^S f_1(z, y_z(\tau, u(\cdot), y), u(\tau)) d\tau \right\}.$$

The admissible controls are defined here and throughout the paper as measurable functions satisfying (1.5). The solutions (admissible trajectories) to the systems considered are assumed to exist and be unique with the use of any admissible control.

Let us introduce the hypothesis that plays a crucial role in our consideration.

Assumption 2.1. With each $z \in W \subset R^{n_1}$, $y \in P \subset R^{n_2}$, the limit

$$(2.3) \quad \lim_{S \rightarrow \infty} \bar{V}(z, S, y) = \bar{V}(z) \quad \forall y \in P$$

exists in the Hausdorff metric, and the evaluation is valid

$$(2.4) \quad \rho(\bar{V}(z, S, y), \bar{V}(z)) \leq \gamma(S) \quad \forall (z, y) \in W \times P,$$

where $\bar{V}(z, S, y)$ is the closure of $V(z, S, y)$; $\bar{V}(z)$ is a convex and compact subset of R^{n_1} , which does not depend on y from P ; $\rho(\cdot, \cdot)$ is the Hausdorff metric; and $\gamma(S) \rightarrow 0$ as $S \rightarrow \infty$.

It can be shown that the convexity of the limit set $\bar{V}(z)$ follows from its existence; however, we do not prove this fact in the paper.

Suppose that Assumption 2.1 is true and consider the differential inclusion

$$(2.5) \quad \dot{z} \in \bar{V}(z), \quad z(0) = z_0.$$

DEFINITION 2.1. We say that the differential inclusion (2.5) approximates z -components of the trajectories of system (1.1),(1.2) if there exists a function $\mu(\epsilon)$ tending to zero as ϵ tends to zero such that, corresponding to any admissible trajectory $\{z_\epsilon(t), y_\epsilon(t)\}$ of system (1.1),(1.2), there exists a solution $z(t)$ to the differential inclusion (2.5), which satisfies the inequality

$$(2.6) \quad \max_{t \in [0, 1]} \|z_\epsilon(t) - z(t)\| \leq \mu(\epsilon).$$

Conversely, given an arbitrary solution $z(t)$ to the differential inclusion (2.5), we can construct an admissible control $u_{z(\cdot)}(t)$, which, being used in system (1.1), (1.2), generates the trajectory $\{z_\epsilon(t), y_\epsilon(t)\}$ satisfying (2.6). The control $u_{z(\cdot)}(t)$ is referred to as z -approximating.

LEMMA 2.1. Suppose that Assumption 2.1 is true and there exist compact sets $D \subset \text{int}W \subset R^{n_1}$, $\Omega \subset P \subset R^{n_2}$ such that the following assumptions are fulfilled.

Assumption 2.2. Each admissible trajectory of the system (1.1),(1.2) satisfies the inclusion $\{z_\epsilon(t), y_\epsilon(t)\} \in D \times \Omega$, for all $t \in [0, 1]$.

Assumption 2.3. The functions f_1, f_2 are continuous, and

$$(2.7) \quad \begin{aligned} \|f_i(z^1, y^1, u) - f_i(z^2, y^2, u)\| &\leq L(\|z^1 - z^2\| + \|y^1 - y^2\|) \\ \forall (z^1, y^1), (z^2, y^2) &\in W \times P, i = 1, 2; \end{aligned}$$

$$(2.8) \quad \|f_1(z, y, u)\| \leq M \quad \forall (z, y, u) \in W \times P \times U,$$

where L and M are constants.

Assumption 2.4. With each $z \in W$ and with any admissible control, the trajectories of the associated system (2.1), which begin in the set Ω , do not leave the set P .

Assumption 2.5. For any $z^1 \in W, z^2 \in W, S > 0$, any admissible control $u(t)$ and any initial values $y \in \Omega$, the solutions to system (2.1) satisfy the inequality

$$(2.9) \quad S^{-1} \left\| \int_0^S f_1(z^1, y_{z^1}(\tau, u(\cdot), y), u(\tau)) d\tau - \int_0^S f_1(z^2, y_{z^2}(\tau, u(\cdot), y), u(\tau)) d\tau \right\| \leq C \|z^1 - z^2\|.$$

Then the differential inclusion (2.5) approximates z -components of the trajectories of system (1.1), (1.2).

The proof of the lemma is given in §5.

Consider the problems of minimization of the functional

$$(2.10) \quad \inf G(z(1))$$

on the set of the admissible trajectories of the singularly perturbed system (1.1), (1.2), on the set of the admissible trajectories of the reduced system (1.3), and on the set of the solutions to the differential inclusion (2.5). We refer to these optimal control problems as singularly perturbed (SP), as reduced (R), and as averaged (A), denoting their optimal values as G_ϵ , G_r , and G_a , respectively.

DEFINITION 2.2. We say that an admissible control $u(t)$ is ϵ -suboptimal in the SP problem if, being used in system (1.1), (1.2), it provides the value of functional (2.10), differing from the optimal value G_ϵ by some function of ϵ that tends to zero as ϵ tends to zero.

COROLLARY 2.1. Assume that the conditions of Lemma 2.1 are satisfied and that the function $G(z)$ is continuous. Then $G_\epsilon \rightarrow G_a$ as $\epsilon \rightarrow 0$. If $z(t)$ is a solution to the A problem, then every z -approximating control $u_{z(\cdot)}(t)$ is ϵ -suboptimal in the SP problem.

Proof. The proof is obvious.

For the R problem to be correctly stated, let us now introduce the following assumption.

Assumption 2.6. With each $(z, u) \in W \times U$, there exists the unique root $\psi(z, u)$ to (1.4), which satisfies the inclusion $\psi(z, u) \in P$. The function $\psi(z, u)$ is continuous in (z, u) on the set $W \times U$ and satisfies the following Lipschitz conditions in z : There exists a constant C_1 such that, for any $z^i, i = 1, 2$ and any $u \in U$,

$$\|\psi(z^1, u) - \psi(z^2, u)\| \leq C_1 \|z^1 - z^2\|.$$

Under Assumptions 2.1 and 2.6, we may write, with any $(z, u) \in W \times U$,

$$(2.11) \quad y_z(t, u, \psi(z, u)) = \psi(z, u)$$

$$(2.12) \quad \begin{aligned} \forall t > 0 &\Rightarrow f_1(z, \psi(z, u), u) \in V(z, S, \psi(z, u)), \\ \forall S > 0 &\Rightarrow f_1(z, \psi(z, u), u) \in \overline{V}(z) \end{aligned}$$

DEFINITION 2.3. The point $\eta \in \overline{V}(z)$ is said to be a stationary regime point if it is presented in the form $\eta = f_1(z, \psi(z, u), u)$ for some $u \in U$. It is said to be a quasi-stationary regime point if it is presented as a convex combination of stationary regime points. The sets of stationary and quasi-stationary regime points are denoted as $V_{st}(z)$ and $\text{conv}V_{st}(z)$, respectively.

DEFINITION 2.4. We say that the SP problem is approximated by the R problem if $G_\epsilon \rightarrow G_r$ as $\epsilon \rightarrow 0$ and if, corresponding to any admissible trajectory $z_\nu(t)$ of the reduced system (1.3) such that $G(z_\nu(1)) \leq G_r + \nu, \nu > 0$, there exists $z_\nu(\cdot)$ -approximating control providing in the SP problem the value of the functional differing from the optimal one by $\nu + \kappa(\epsilon)$, where $\kappa(\epsilon) \rightarrow 0$ as $\epsilon \rightarrow 0$.

LEMMA 2.2. Suppose that the function $G(z)$ is continuous, the set U is compact, and Assumptions 2.1–2.6 are true. Then the SP problem is approximated by the R problem if and only if there exists a solution $z(t)$ to the A problem such that, for

almost all $t \in [0, 1]$, the velocity vector $\dot{z}(t)$ is a quasi-stationary regime point of the set $\bar{V}(z(t))$.

Proof. Note first that, according to (2.11), each trajectory of the reduced system is at the same time a solution to the differential inclusion (2.4). Consequently, $G_a \leq G_r$. From Corollary 2.1, it follows that the SP problem is approximated by the R problem if and only if

$$(2.12) \quad G_a = G_r.$$

Suppose that $z(t)$ is a solution to the A problem satisfying the relation $\dot{z}(t) \in \text{conv}V_{st}(z(t))$ for almost all $t \in [0, 1]$. From Filippov's result [9, Thm. 3] it follows that this solution can be presented as the limit $z(t) = \lim_{n \rightarrow \infty} z_n(t)$ in the uniform metric, where $z_n(t)$ are solutions of the differential inclusion $\dot{z}_n(t) \in V_{st}(z_n(t))$. According to Filippov's lemma [8], there exist admissible controls $u_n(t)$ such that $\dot{z}_n(t) = f_1(z_n(t), \psi(z_n(t), u_n(t)), u_n(t))$ for almost all $t \in [0, 1]$; that is, $z_n(t)$ are admissible trajectories of the reduced system and $G(z_n(1)) \geq G_r$. Passing to the limit as n tends to infinity, we obtain that $G_a \stackrel{\text{def}}{=} G(z(1)) \geq G_r$. As we remarked above, the converse inequality is always true, and thus equality (2.12) is valid. Conversely, the fulfillment of this equality means that a solution to problem $\min\{G(z(1)) | \dot{z} \in \text{conv}V_{st}(z), z(0) = 0\}$ is also a solution to the A problem. \square

General conditions for Assumption 2.1 to be true were obtained in Gaitsgory [11]. We now consider a class of singularly perturbed optimal control problems satisfying these conditions and admitting a description of the limit set $\bar{V}(z)$ in terms of stationary, quasi-stationary, and also periodic regime (see Definition 3.1) points.

3. Structure of the limit sets. In this section, we study properties of the associated system, with z entering only as a vector of constant parameters. For brevity, we omit it in our notations, writing $f_1(y, u)$, $f_2(y, u)$ instead of $f_1(z, y, u)$, $f_2(z, y, u)$. System (2.1) is then rewritten in the form

$$(3.1) \quad \dot{y} = f_2(y, u),$$

and set (2.3) in the form

$$V(S, y) = \bigcup \{S^{-1} \int_0^S f_1(y(\tau, u(\cdot), y), u(\tau)) d\tau\},$$

where, here and in every place that follows, $y(\tau, u(\cdot), y)$ is the solution to (3.1) obtained with the admissible control $u(\tau)$ and with the initial values y . The union is taken over the all admissible controls. Denote by $Y(S, y)$ the reachability set of system (3.1)

$$Y(S, y) \stackrel{\text{def}}{=} \bigcup \{y(S, u(\cdot), y)\}, Y(S, Q) \stackrel{\text{def}}{=} \bigcup \{Y(S, y)\},$$

where the unions are taken over the all admissible controls and over the initial values y from a set $Q \subset R^{n_2}$, respectively. Note that

$$(3.2) \quad Y(S + S') = Y(S, Y(S', y)).$$

THEOREM 3.1. *Suppose that the functions $f_1(y, u)$, $f_2(y, u)$ are continuous and satisfy Lipschitz conditions with respect to y in a sufficiently large domain. Suppose that the set U is compact and that there exists a compact set P such that the following assumptions are true.*

Assumption 3.1. The solutions to the system (3.1) obtained with any initial values $y^i \in P, i = 1, 2$ and with any admissible control $u(\cdot)$ satisfy the inequality

$$(3.3) \quad \|y(\tau, u(\cdot), y^1) - y(\tau, u(\cdot), y^2)\| \leq \xi(\tau) \|y^1 - y^2\|$$

with the function $\xi(\tau)$ satisfying the relations

$$(3.4) \quad \int_0^\infty \xi(\tau) d\tau < \infty, \quad \lim_{\tau \rightarrow \infty} \xi(\tau) = 0.$$

Assumption 3.2. There exists a point $y^* \in P$ such that, with any admissible control $u(\tau)$, the inclusion $y(\tau, u(\cdot), y^*) \in P$, for all $\tau > 0$ holds true.

Then: (i) Assumption 2.1 is fulfilled; that is, the limit of the closure $\lim_{S \rightarrow \infty} \bar{V}(S, y) \stackrel{\text{def}}{=} \bar{V}$ exists in the Hausdorff metric. This limit is a convex and closed subset of R^{n_1} , and the following evaluation is valid:

$$(3.5) \quad \rho(\bar{V}(S, y), \bar{V}) \leq FS^{-0.5}, \quad S \geq S_0, y \in P,$$

where F and S_0 are some positive constants; and

(ii) The limit of the closure $\lim_{S \rightarrow \infty} \bar{Y}(S, y) \stackrel{\text{def}}{=} \bar{Y}$ exists in the Hausdorff metric. This limit is invariant with respect to system (3.1) as follows:

$$(3.6) \quad Y(S, \bar{Y}) \subset \bar{Y} \subset P, \bar{Y}(S, \bar{Y}) = \bar{Y} \quad \forall S > 0$$

The following evaluation is valid:

$$(3.7) \quad \rho(\bar{Y}(S, y), \bar{Y}) \leq 2C\xi(S), \quad C \stackrel{\text{def}}{=} \max\{\|y^1 - y^2\| | y_i \in P, i = 1, 2\}.$$

Proof. Statement (i) follows from a more general result in Gaitsgory [11]. To prove (ii), let us first note that from (3.3) it follows that

$$(3.8) \quad \rho(\bar{Y}(S, Q^1), \bar{Y}(S, Q^2)) \leq \xi(S) \rho(Q^1, Q^2),$$

where Q^1, Q^2 are arbitrary subsets of P . In particular, with $Q^1 = y^*, Q^2 = Q$, we have that

$$(3.9) \quad \rho(\bar{Y}(S, y^*), \bar{Y}(S, Q)) \leq \rho(y^*, Q) \xi(S) \leq C\xi(S).$$

According to Assumption 3.2, $Y(S', y^*) \subset P$, for all $S' > 0$. Consequently, if we take $Y(S', y^*)$ as Q in (3.9), then $\rho(\bar{Y}(S, y^*), \bar{Y}(S, Y(S', y^*))) \leq C\xi(S)$. In view of (3.2), from here it follows that $\rho(\bar{Y}(S, y^*), \bar{Y}(S + S', y^*)) \leq C\xi(S)$, which means that, with an arbitrary way of S tending to infinity, the sequence $\bar{Y}(S, y^*)$ is fundamental in the space of the closed subsets of the compact set P provided with the Hausdorff metric. By Blaschke's theorem (see Hadwiger [17]), this metric space is compact, and hence the following limit exists: $\lim_{S \rightarrow \infty} \bar{Y}(S, y^*) \stackrel{\text{def}}{=} \bar{Y} \subset P$. To estimate the rate of convergence, let us write

$$\begin{aligned} \rho(\bar{Y}(S, y^*), \bar{Y}) &\leq \rho(\bar{Y}(S, y^*), \bar{Y}(S + S', y^*)) + \rho(\bar{Y}(S + S', y^*), \bar{Y}) \\ &\leq C\xi(S) + \rho(\bar{Y}(S + S', y^*), \bar{Y}); \end{aligned}$$

hence, with S' tending to infinity, we obtain that

$$\begin{aligned}\rho(\bar{Y}(S, y^*), \bar{Y}) &\leq C\xi(S) \Rightarrow \rho(\bar{Y}(S, y), \bar{Y}) \leq \rho(\bar{Y}(S, y), \bar{Y}(S, y^*)) + \rho(\bar{Y}(S, y^*), \bar{Y}) \\ &\leq \|y - y^*\|\xi(S) + C\xi(S) \leq 2C\xi(S) \quad \forall y \in P.\end{aligned}$$

We may further write that

$$\begin{aligned}\rho(\bar{Y}(S, \bar{Y}), \bar{Y}) &\leq \rho(\bar{Y}(S, \bar{Y}), \bar{Y}(S + S', y^*)) + \rho(\bar{Y}(S + S', y^*), \bar{Y}) \\ &\leq \xi(S)\rho(\bar{Y}, Y(S', y^*)) + C\xi(S + S') \leq C(\xi(S)\xi(S') + \xi(S + S')).\end{aligned}$$

Passing to the limit with $S' \rightarrow \infty$, we obtain (3.6). \square

Now we proceed to a characterization of the structure of the limit set \bar{Y} . We begin with the following statement.

LEMMA 3.1. *Suppose that the assumptions of Theorem 3.1 are true. Then*

(i) *Corresponding to any $T \geq T_0$, where T_0 is a positive number, and to any admissible control $u(t)$ defined on the interval $[0, T]$ there exists a unique solution $y(t)$ to system (3.1) satisfying the periodicity conditions*

$$(3.10) \quad y(0) = y(T) \in P.$$

This solution is contained completely in the limit reachability set \bar{Y} ;

(ii) *Corresponding to any $u \in U$, there exists the unique root $y = \psi(u) \in P$ of the equation*

$$(3.11) \quad f_2(y, u) = 0;$$

(iii) *The function $\psi(u)$ is continuous on U .*

Proof. Note that, if there exists an admissible trajectory $y(t)$ of system (3.1) that satisfies (3.10), it belongs to \bar{Y} . It is implied by that, if we complete the definition of $y(t)$ on the interval $[T, \infty]$, considering it as the periodic function, then $y(t) = y(t + kT) \in \bar{Y}(t + kT, y(0))$ with an arbitrary natural k and any $t \in [0, T]$. Passing to the limit with $k \rightarrow \infty$, we obtain the inclusion $y(t) \in \bar{Y}$.

Suppose that a constant T_0 is chosen in such a way that $\xi(T) \leq \delta < 1$ with any $T \geq T_0$. Define the operator $A_{u(\cdot)}(y) \stackrel{\text{def}}{=} y(T, u(\cdot), y)$ with $y \in \bar{Y}$. As it follows from (3.6), this operator reflects \bar{Y} into \bar{Y} . On the other hand, it belongs to the class of the contractive operators, since $\|A_{u(\cdot)}(y^1) - A_{u(\cdot)}(y^2)\| \leq \delta\|y^1 - y^2\|$. Consequently, there exists the unique fixed point of the operator $A_{u(\cdot)}$, which defines the solution satisfying (3.10). Thus (i) is proved.

Let u be an arbitrary vector from U . Define the constant control $u(t) = u$. In accordance with (i), with any $T \geq T_0$ corresponding to this control, there exists the T -periodic solution to system (3.1). On the basis of Assumption 3.1, it is easy to verify that all these solutions coincide, which may take place only if they do not depend on the time. Thus, corresponding to the constant control $u(t) = u$, there exists the unique constant solution to system (3.1) defined as the unique root

$$(3.12) \quad y = \psi(u) \in \bar{Y}$$

of (3.11). To prove that the function $\psi(u)$ is continuous, suppose that $u_n \rightarrow u$ as $n \rightarrow \infty$, where $u_n \in U$ and, consequently, $u \in U$. Denote by y^0 an arbitrary partial limit: $\psi(u_{n'}) \stackrel{\text{def}}{=} y_{n'} \rightarrow y^0$, as $n' \rightarrow \infty, \{n'\} \subset \{n\}$. Such limits exist, since the

inclusion (3.12) is valid and the set \bar{Y} is compact. The continuity of the function $\psi(u)$ is established if we show that $y^0 = \psi(u)$. However, this is implied by the relations

$$0 = f_2(y_{n'}, u_{n'}) \rightarrow f_2(y^0, u) \Rightarrow f_2(y^0, u) = 0$$

and by the fact that the root of (3.11) is unique. \square

A stationary regime point η was defined in §2, presented in the form

$$\eta = f_1(\psi(u), u) \in \bar{V}.$$

Note that according to statement (iii) of Lemma 3.1, the set V_{st} of such points is a compact subset of \bar{V} . We introduce the following definition.

DEFINITION 3.1. The point $\eta \in \bar{V}$ is said to be a periodic regime point if there exists a positive number $T > 0$, an admissible control $u(t)$ defined on the interval $[0, T]$, and the solution $y(t)$ to system (3.1) satisfying the periodicity conditions (3.10) such that $\eta = T^{-1} \int_0^T f_1(y(t), u(t))dt$. The set of the periodic regime points is denoted as V_p .

THEOREM 3.2. Let the assumptions of Theorem 3.1 be true. Then

(i) The limit set \bar{V} is equal to the closure of the set of the periodic regime points $\bar{V} = \bar{V}_p$;

(ii) The set \bar{V} consists of the quasi-stationary regime points only: $\bar{V} = \text{conv} V_{st}$ if and only if, with any $\lambda \in R^{n_1}$,

$$(3.13) \quad h_{st}(\lambda) = h_p(\lambda).$$

Here $h_{st}(\lambda)$ is the optimal value of the "steady state" optimization problem

$$(3.14) \quad h_{st}(\lambda) \stackrel{\text{def}}{=} \min\{\lambda^T f_1(\psi(u), u) | u \in U\},$$

and $h_p(\lambda)$ is the optimal value of the periodic optimization problem

$$(3.15) \quad h_p(\lambda) \stackrel{\text{def}}{=} \inf\{T^{-1} \int_0^T \lambda^T f_1(y(t), u(t))dt\},$$

where \inf is sought over the length T of the time interval, over the admissible controls defined on $[0, T]$, and over the corresponding solutions to system (3.1), which satisfy the periodicity conditions (3.10).

Proof. Let $\eta \in \bar{V}$. By virtue of (3.5), with any $T \geq S_0$ and any $\nu \geq 0$, there exists a vector η' such that $\|\eta - \eta'\| \leq FT^{-0.5} + \nu$; $\eta' \in V(T, y)$, $y \in P$. By the definition, the latter inclusion means that there exists an admissible control $u(t)$ such that $\eta' = T^{-1} \int_0^T f_1(y(t, u(\cdot), y), u(t))dt$. From statement (i) of Lemma 3.1, it follows that, if $T \geq T^0$, then, corresponding to the control $u(t)$, there exists the unique solution $y(t)$ to system (3.1) satisfying the periodicity conditions (3.10). Define the periodic regime point $\eta'' \stackrel{\text{def}}{=} T^{-1} \int_0^T f_1(y(t), u(t))dt$. Using (3.3), we may write

$$\begin{aligned} \|\eta' - \eta''\| &\leq T^{-1} \int_0^T \|f_1(y(t, u(\cdot), y), u(t)) - f_1(y(t), u(t))\|dt \\ &\leq T^{-1} L \int_0^T \|y(t, u(\cdot), y) - y(t)\|dt \leq T^{-1} L \|y - y(0)\| \int_0^\infty \xi(t)dt \leq T^{-1} q, \end{aligned}$$

where L is a Lipschitz constant and $q = LC \int_0^\infty \xi(t)dt$, C is defined in (3.7). Consequently,

$$\text{dist}(\eta, V_p) \stackrel{\text{def}}{=} \inf\{\|\eta - \zeta\| \mid \zeta \in V_p\} \leq \|\eta - \eta'\| + \|\eta' - \eta''\| \leq FT^{-0.5} + \nu + T^{-1}q.$$

Considering that η is an arbitrary vector from \bar{V} and ν is an arbitrary positive number, and passing to the limit with T tending to infinity, we obtain that $\sup\{\text{dist}(\eta, V_p) \mid \eta \in \bar{V}\} = 0$, which implies that $\rho(\bar{V}, V_p) = 0$, since $V_p \subset \bar{V}$. This proves (i). To prove (ii), let us remark that the optimal values (3.14), (3.15) can also be presented in the forms $h_{st}(\lambda) = \min\{\lambda^T \eta \mid \eta \in V_{st}\}$, $h_p(\lambda) = \inf\{\lambda^T \eta \mid \eta \in V_p\}$. These representations and (i) permit us to write $h_{st}(\lambda) = \min\{\lambda^T \eta \mid \eta \in \text{conv}V_{st}\}$, $h_p(\lambda) = \min\{\lambda^T \eta \mid \eta \in \bar{V}\}$. It means that $h_{st}(\lambda)$ and $h_p(\lambda)$ are the supporting functions to convex and compact sets $\text{conv}V_{st}$ and \bar{V} . The equality of these functions is equivalent to the equality of the corresponding sets. \square

Note that, in a general case, the relation $h_p(\lambda) \leq h_{st}(\lambda)$ is valid, and different tests developed in periodic optimization theory (see Bailey and Horn [2], Cirilin, Balakirev, and Dudnikov [4], Gilbert [15], Guardabassi, Locatelli, and Rinaldi [16], and others) can be used to establish whether this relation takes the form of a strict inequality or the form of equality (3.13). Thus it can be used to verify the validity of the representation $\bar{V} = \text{conv}V_{st}$. Let us now consider some other sufficient conditions for this representation to be true.

DEFINITION 3.2. (Filippov [10], Gelig, Leonov, and Iakubovich [14]). The set $Y_{st} \stackrel{\text{def}}{=} \{y \mid y = \psi(u), u \in U\}$ is said to be stable in Lyapunov sense if, corresponding to any $\mu > 0$, there exists $\delta > 0$ such that, with any admissible control $u(t)$ from the inequality $\text{dist}(y, Y_{st}) < \delta$, it follows that $\text{dist}(y(t, u(\cdot), y), Y_{st}) < \mu$, for all $t > 0$.

PROPOSITION 3.1. Suppose, in addition to the assumptions of Theorem 3.1, that the set Y_{st} is stable in Lyapunov sense and that $f_1(y, u) \equiv f_1(y)$. Then $\bar{Y} = Y_{st}$ and $\bar{V} = \text{conv}V_{st}$.

Proof. According to (3.12), $Y_{st} \subset \bar{Y}$. On the other hand, from Lyapunov stability of Y_{st} , it follows that, if $y \in Y_{st}$, then $Y(S, y) \subset Y_{st}$ for any $S > 0$. Hence, by statement (ii) of Theorem 3.1, $\bar{Y} \subset Y_{st}$. Thus $\bar{Y} = Y_{st}$.

By virtue of (3.6), we may write, with any admissible control and any initial values

$$y \in Y_{st}, y(S, u(\cdot), y) \in Y(S, y) \subset \bar{Y} = Y_{st}.$$

It follows that $S^{-1} \int_0^S f_1(y(t, u(\cdot), y))dt \in \text{conv}f_1(Y_{st}) \stackrel{\text{def}}{=} \text{conv}V_{st}$. Consequently, $V(S, y) \subset \text{conv}V_{st}$ and $\bar{V} \subset \text{conv}V_{st}$. Since the converse is obvious, it completes the proof. \square

PROPOSITION 3.2. Under the conditions of Theorem 3.1, let the set

$$(3.16) \quad f(\bar{Y}, U) \stackrel{\text{def}}{=} \{\eta \mid \eta = f(y, u), (y, u) \in \bar{Y} \times U\}$$

be convex, where $f(y, u) \stackrel{\text{def}}{=} \{f_1(y, u), f_2(y, u)\}$, then $\bar{V} = V_{st}$.

Proof. Define the set $P(S, y) \stackrel{\text{def}}{=} \bigcup \{S^{-1} \int_0^S f(y(t, u(\cdot), y), u(t))dt\}$, where the union is taken over all admissible controls. Note that, by this definition, $V(S, y) = \{\eta_1 \mid (\eta_1, \eta_2) \in P(S, y)\}$. In accordance with (3.6) and the assumption about the convexity of the set (3.16), we may write $P(S, y) \subset \text{conv}f(\bar{Y}, U) = f(\bar{Y}, U)$, for all $y \in \bar{Y}$. On the other hand, by virtue of (3.1), $\|S^{-1} \int_0^S f_2(y(t, u(\cdot), y), u(t))dt\| =$

$\|S^{-1} \int_0^S \dot{y}(t, u(\cdot), y) dt\| = S^{-1} \|y(S, u(\cdot), y) - y\| \leq CS^{-1}$, where C is defined in (3.7). Consequently,

$$P(S, y) \subset \{(\eta_1, \eta_2) | (\eta_1, \eta_2) \in f(\bar{Y}, U), \|\eta_2\| \leq CS^{-1}\}.$$

Denoting the right-hand side in the last inclusion by $Q(S, y)$, we can easily verify that $\rho(Q(S, y), Q) \stackrel{\text{def}}{=} \mu(S)$ tends to zero as S tends to infinity, where

$$Q \stackrel{\text{def}}{=} \{(\eta_1, \eta_2) | (\eta_1, \eta_2) \in f(\bar{Y}, U), \eta_2 = 0\} = \{(\eta_1, 0) | (\eta_1, 0) \in f(\bar{Y}, U)\}.$$

Hence $P(S, y) \subset Q + \mu(S)\bar{B}^{n_1+n_2}$ and

$$V(S, y) \subset \{\eta_1 | (\eta_1, \eta_2) \in Q + \mu(S)\bar{B}^{n_1+n_2}\} = \{\eta_1 | (\eta_1, \eta_2) \in Q\} + \mu(S)\bar{B}^{n_1},$$

where $\bar{B}^{n_1+n_2}$, \bar{B}^{n_1} are the closed balls with the centers in the origin and with the unit radii in $R^{n_1+n_2}$ and R^{n_1} , respectively. Noting now that

$$\{\eta_1 | (\eta_1, \eta_2) \in Q\} \stackrel{\text{def}}{=} \{\eta_1 | \eta_1 = f_1(y, u), 0 = f_2(y, u), (y, u) \in \bar{Y} \times U\} = V_{st},$$

we obtain the inclusion $V(S, y) \subset V_{st} + \mu(S)\bar{B}^{n_1}$, which implies that $\bar{V} \subset V_{st}$. Since the converse is obvious, the proof is complete. \square

Consider in conclusion some examples.

Example 3.1. Suppose that system (3.1) is linear and is presented in the form

$$(3.17) \quad \dot{y} = A_2 y + B_2 u + F_2,$$

where A_2, B_2, F_2 are matrices of the corresponding dimensions, and the eigenvalues of the matrix A_2 have negative real parts. Under this supposition, inequality (3.3) is fulfilled with any y^1, y^2 from R^{n_2} and with $\xi(t) = e^{-\alpha t}$, where α is a positive number defined by the absolute values of the real parts of the matrix A_2 eigenvalues. If the set U is compact, from the stability of the matrix A_2 , it follows that the admissible trajectories of system (3.17), which begin at point $y^* = 0$, do not leave some closed ball \bar{B} with the center in the origin.

Thus the linearity and stability of system (3.1) and the compactness of the set U provide the fulfillment of the conditions of Theorem 3.1, with P defined as an arbitrary compact set containing \bar{B} .

Example 3.2. Let $n_1 = 1, n_2 = 2, m = 1, U = \{u | |u| \leq 1\}, y = (y_1, y_2), f_1(y, u) = u^2 - y_1^2, f_2(y, u) = A_2 y + D_2 u$, with

$$A_2 = \begin{pmatrix} 0 & 1 \\ -\omega^2 & -k \end{pmatrix}, \quad D_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad k > 0.$$

The eigenvalues of matrix A_2 have negative real parts, and thus the conditions of Theorem 3.1 are satisfied. Let us show that $\bar{V} \neq \text{conv} V_{st}$ if

$$(3.18) \quad \omega > 1, \omega k < 1.$$

In the scalar case under consideration, equality (3.13) should be verified only for the following two values of λ : $\lambda = 1$ and $\lambda = -1$. Considering (3.18), we obtain, with $\lambda = 1$, that

$$h_{st}(1) = \min\{u^2 - y_1^2 | y_2 = 0, -y_1\omega^2 + u = 0, |u| \leq 1\} = \min\{u^2 \left(1 - \frac{1}{\omega^2}\right) | |u| \leq 1\} = 0.$$

On the other hand, the periodic solution corresponding to the control

$$(3.19) \quad u(t) = \cos \omega t$$

is of the form $y_1(t) = (1/\omega k) \sin \omega t$, $y_2(t) = (1/k) \cos \omega t$, and

$$h_p(1) \leq (\omega/2\pi) \int_0^{2\pi/\omega} (\cos^2 \omega t - (1/\omega^2 k^2) \sin^2 \omega t) dt = \frac{1}{2}(1 - (1/\omega^2 k^2)) < 0.$$

With $\lambda = -1$, we obtain that $h_{st}(-1) = \min\{-u^2(1 - 1/\omega^2)||u| \leq 1\} = 1/\omega^2 - 1$ and $h_p(-1) = -1$, where the minimum in the periodic optimization problem is achieved on the sliding regime ($u = \pm 1, y_1 = y_2 = 0$).

Thus, in this example,

$$\text{conv} V_{st} = \left[0, 1 - \frac{1}{\omega^2}\right] \subset \left[\frac{1}{2} \left(1 - \frac{1}{\omega^2 k^2}\right), 1\right] \subset \bar{V}.$$

Example 3.3. Suppose that $n_2 = 1, m = 2$, and system (3.1) is written in the form

$$(3.20) \quad \dot{y} = (a - \alpha y)u_1 + (b - \beta y)u_2,$$

where $\alpha > 0, \beta > 0, a\alpha^{-1} < b\beta^{-1}$, $U = \{u = (u_1, u_2) | u_1 + u_2 = 1; u_i \geq 0, i = 1, 2\}$. With any initial values y^1, y^2 from R^1 and any admissible control $u(t) = (u_1(t), u_2(t))$, difference between the solutions $y(t, u(\cdot), y^1) - y(t, u(\cdot), y^2) \stackrel{\text{def}}{=} \delta(t)$ satisfies the equation

$$\dot{\delta} = -(\alpha u_1(t) + \beta u_2(t))\delta, \delta(0) = y^1 - y^2.$$

Consequently, $\delta(t) = (y^1 - y^2) \exp(-\int_0^t (\alpha u_1(\tau) + \beta u_2(\tau)) d\tau)$ and

$$|\delta(t)| \leq |y^1 - y^2| e^{-\gamma t}, \gamma \stackrel{\text{def}}{=} \min\{\alpha u_1 + \beta u_2 | u \in U\} = \min(\alpha, \beta).$$

Thus inequality (3.3) is true with $\xi(t) = e^{-\gamma t}$. It is also easy to check the following.

(i) The trajectories of system (3.20), which begin in the interval $[a\alpha^{-1}, b\beta^{-1}]$, do not leave this interval, and, consequently, Assumptions 3.1 and 3.2 are true with any compact set P containing $[a\alpha^{-1}, b\beta^{-1}]$.

(ii) Interval $[a\alpha^{-1}, b\beta^{-1}]$ coincides with set Y_{st} , and this set is stable in the Lyapunov sense. Thus, according to Proposition 3.1, $\bar{Y} = Y_{st}$, and, if $f_1(y, u) \equiv f_1(y)$, then $\bar{V} = \text{conv} V_{st} = \text{conv} f_1(Y_{st})$.

Example 3.4. Suppose that system (3.1) is written in the form (3.17) and the eigenvalues of the matrix A_2 have negative real parts. Suppose also that U is a convex compact set and the function $f_1(y, u)$ is linear: $f_1(y, u) = A_1 y + B_1 u + F_1$, where A_1, B_1, F_1 are matrices. Then, from the convexity of the limit reachability set \bar{Y} , which is implied by the convexity of the reachability set $Y(S, y)$, it follows that set (3.16) is convex, and, by virtue of Proposition 3.2,

$$(3.21). \quad \bar{V} = V_{st} = \{\eta | \eta = \tilde{B}u + \tilde{F}, u \in U\}, \tilde{B} \stackrel{\text{def}}{=} B_1 - A_1 A_2^{-1} B_2, \tilde{F} \stackrel{\text{def}}{=} F_1 - A_1 A_2^{-1} F_2.$$

4. Final results. In this section we use previous results to describe a class of singularly perturbed control systems characterized by the fulfillment of Assumption 3.1, which we reformulate below, considering the dependence of the associated system on z .

THEOREM 4.1. *Suppose that the set U is compact and that there exist compact sets $D \subset \text{int}W \subset R^{n_1}$, $\Omega \subset P \subset R^{n_2}$ such that Assumptions 2.2–2.4 are fulfilled. Assume also that the following is true.*

Assumption 4.1. With each $z \in W$, the trajectories of the associated system (2.1) obtained with any initial values $y^i \in P$, $i = 1, 2$ and with any admissible control $u(\cdot)$ satisfy the inequality

$$(4.1) \quad \|y_z(\tau, u(\cdot), y^1) - y_z(\tau, u(\cdot), y^2)\| \leq \xi(\tau) \|y^1 - y^2\|,$$

where $\xi(\tau)$ does not depend on z and satisfies relation (3.4).

Then (i) Assumption 2.1 is satisfied and estimate (2.4) takes the form

$$(4.2) \quad \rho(\bar{V}(z, S, y), \bar{V}(z)) \leq LS^{-0.5} \quad \forall (z, y) \in W \times P;$$

(ii) With each $z \in W$, the limit reachability set $\bar{Y}(z)$ of the associated system exists, and the rate of the convergence to this limit is estimated by the inequality

$$(4.3) \quad \rho(\bar{Y}(z, S, y), \bar{Y}(z)) \leq 2C\xi(S);$$

(iii) z -components of the trajectories of system (1.1), (1.2) are approximated by the trajectories of the differential inclusion (2.5).

Proof. The existence and the convexity of the limit set $\bar{V}(z)$ and the existence of the limit set $\bar{Y}(z)$ are implied by Theorem 3.1. The uniformity of estimates (4.2), (4.3) follows from the fact that the function $\xi(\tau)$ in (4.1) and the set P are supposed to be independent on $z \in W$ (the constant C is defined in (3.7), and the constant L can be written explicitly using estimates obtained in the proof of Theorem 4.1 in Gaitsgory [11]).

In accordance with Lemma 2.1, to prove statement (iii), it remains to verify Assumption 2.5.

LEMMA 4.1. *If $\tilde{z}(t)$ is an arbitrary continuous function such that $\tilde{z}(t) \in W$, for all $t \in [0, S]$, $u(t)$ is an arbitrary admissible control, and $\tilde{y}(t)$ is the absolutely continuous function satisfying the following relations: $\tilde{y}(t) \in \Omega$, for all $t \in [0, S]$ and*

$$(4.4) \quad \dot{\tilde{y}} = f_2(\tilde{z}(t), \tilde{y}(t), u(t)), \tilde{y}(0) = y \in \Omega,$$

then

$$(4.5) \quad \max_{t \in [0, S]} \|\tilde{y}(t) - y_z(t, u(\cdot), y)\| \leq C_1 \max_{t \in [0, S]} \|\tilde{z}(t) - z\| \quad \forall z \in W,$$

where C_1 is a constant.

Proof of the lemma. Let σ be a positive number such that $\xi(\sigma) = \delta < 1$. Divide the interval $[0, S]$ by the points: $\tau_l \stackrel{\text{def}}{=} l\sigma$, $l = 0, 1, \dots, N \stackrel{\text{def}}{=} [S\sigma^{-1}]$, $\tau_{N+1} \stackrel{\text{def}}{=} S$. On each interval $[\tau_l, \tau_{l+1}]$, define the function $y_z^l(\tau)$ as the solution to system (2.1) obtained with the control $u(\tau)$ and with the initial values $y_z^l(\tau_l) = \tilde{y}(\tau_l)$. Obviously, this solution satisfies the equation

$$(4.6) \quad y_z^l(t) = \tilde{y}(\tau_l) + \int_{\tau_l}^t f_2(z, y_z^l(\tau), u(\tau)) d\tau.$$

Since the function $\tilde{y}(\tau)$ also satisfies the equality

$$(4.7) \quad \tilde{y}(t) = \tilde{y}(\tau_l) + \int_{\tau_l}^t f_2(\tilde{z}(\tau), \tilde{y}(\tau), u(\tau)) d\tau,$$

we may subtract it from (4.6) and obtain, using (2.7), that $\|y_z^l(t) - \tilde{y}(t)\| \leq L\sigma\Delta + L \int_{\tau_l}^t \|y_z^l(\tau) - \tilde{y}(\tau)\| d\tau$, where $\Delta \stackrel{\text{def}}{=} \max_{t \in [0, S]} \|\tilde{z}(t) - z\|$. By virtue of the Gronwall–Bellman lemma, from here it follows that $\|y_z^l(\tau_{l+1}) - \tilde{y}(\tau_{l+1})\| \leq L\sigma\Delta e^{L\sigma}$, $l = 0, 1, \dots, N-1$, which permits us to write

$$\begin{aligned} \|y_z(\tau_{l+1}, u(\cdot), y) - \tilde{y}(\tau_{l+1})\| &\leq \|y_z(\tau_{l+1}, u(\cdot), y) - y_z^l(\tau_{l+1})\| + L\sigma\Delta e^{L\sigma} \\ &\leq \|y_z(\tau_l, u(\cdot), y) - y_z^l(\tau_l)\| \xi(\tau_{l+1} - \tau_l) + L\sigma\Delta e^{L\sigma} \\ &= \|y_z(\tau_l, u(\cdot), y) - \tilde{y}(\tau_l)\| \delta + L\sigma\Delta e^{L\sigma} \Rightarrow \\ &\|y_z(\tau_l, u(\cdot), y) - \tilde{y}(\tau_l)\| \leq \Delta \frac{L\sigma e^{L\sigma}}{1 - \delta}, \quad l = 0, 1, \dots, N. \end{aligned}$$

The solution $y_z(\tau, u(\cdot), y)$ to the associated system satisfies the same equation as (4.6), with the replacement of $\tilde{y}(\tau_l)$ by $y_z(\tau_l, u(\cdot), y)$. Subtracting it from (4.7), we have that

$$\|y_z(t, u(\cdot), y) - \tilde{y}(t)\| \leq \Delta L\sigma \left(\frac{e^{L\sigma}}{1 - \delta} + 1 \right) + L \int_{\tau_l}^t \|y_z(\tau, u(\cdot), y) - \tilde{y}(\tau)\| d\tau.$$

Again applying the Gronwall–Bellman lemma, we obtain that

$$\|y_z(t, u(\cdot), y) - \tilde{y}(t)\| \leq \Delta L\sigma e^{L\sigma} \left(\frac{e^{L\sigma}}{1 - \delta} + 1 \right) \quad \forall t \in [\tau_l, \tau_{l+1}], \quad l = 0, 1, \dots, N,$$

which proves the lemma with $C_1 = L\sigma e^{L\sigma} (e^{L\sigma}/(1 - \delta) + 1)$.

If we now take $\tilde{z}(t) \stackrel{\text{def}}{=} z^1 \in W$, $z = z^2 \in W$, from (4.5), it follows that

$$(4.8) \quad \|y_{z^1}(t, u(\cdot), y) - y_{z^2}(t, u(\cdot), y)\| \leq C_1 \|z^1 - z^2\| \quad \forall t > 0.$$

This inequality and the Lipschitz conditions (2.7) imply Assumption 2.5, and thus the proof is complete. \square

Example 4.1. Suppose that

$$(4.9) \quad f_2(z, y, u) = A_2(z)y + B_2(z)u + F_2(z),$$

where $A_2(z)$, $B_2(z)$, $F_2(z)$ are matrices functions with

$$(4.10) \quad \|e^{A_2(z)t}\| \leq e^{-\alpha t}, \alpha > 0 \quad \forall z \in W.$$

This supposition provides the fulfillment of Assumptions 4.1 and 2.4 if P is chosen as a compact set containing the trajectories of the associated system that begin in Ω (see Example 3.1).

THEOREM 4.2. *Under the conditions of Theorem 4.1, the SP problem is approximated by the R problem if and only if there exists a solution $z(t)$ to the A problem such that, for almost all $t \in [0, 1]$, the velocity vector $\dot{z}(t)$ is a quasi-stationary regime point of the set $\bar{V}(z(t))$.*

Proof. According to Lemma 2.2, to prove the theorem, we should verify the fulfillment of Assumption 2.6. The uniqueness of the root $\psi(z, u)$ of (1.4) follows from statement (ii) of Lemma 3.1. The continuity of the function $\psi(z, u)$ is established on the base of the uniqueness just in the same way as in statement (iii) of the mentioned lemma. If we take a constant control $u(t) = u$, then, in accordance with (4.1),

$y_z(t, u, y) \rightarrow \psi(z, u)$ as t tends to infinity. Taking such control and passing to the limit in (4.8), we obtain that $\|\psi(z^1, u) - \psi(z^2, u)\| \leq C_1 \|z^1 - z^2\|$. \square

Example 4.2. Consider the optimal control problem

$$(4.11) \quad \min \left\{ \int_0^1 (u^2 - y^2) dt \mid \epsilon^2 \ddot{y} + \epsilon k \dot{y} + \omega^2 y = u, u \in U \right\}, U \stackrel{\text{def}}{=} \{u \mid |u| \leq 1\} \subset \mathbb{R}^1.$$

After the standard replacement of the variables $y_1 \stackrel{\text{def}}{=} y$, $y_2 \stackrel{\text{def}}{=} \dot{y}$, $\dot{z} = u^2 - y^2$, the problem takes the form of the SP one, with f_1, f_2 defined as in Example 3.2 and with $G(z) = z$. The R problem takes here the form

$$(4.12) \quad \min \{z(1) \mid \dot{z} = u^2(1 - \frac{1}{\omega^2}), u \in U\}.$$

Under conditions (3.18), the optimal value of this problem is equal to zero. At the same time, the control (3.19) delivers to the functional the value $\frac{1}{2}(1 - 1/\omega^2 k^2) + O(\epsilon)$. Thus problem (4.12) does not approximate (4.11). In terms of the described formalism, it is connected with that the boundary points of the limit set \bar{V} are not quasi-stationary regime (see Example 3.2). On the other hand, we may note that the gap between the optimal values in (4.11) and (4.12) admits a clear mechanical interpretation. Condition (3.18) postulates a relative smallness of the friction coefficient k comparatively with the proper frequency ω , which makes it possible to diminish the value of the functional via the resonance oscillations of the variables. The example was proposed as such in an interpretation by Pervozvansky.

In a general case, some boundary points of the set $\bar{V}(z)$ may belong to the set of the quasi-stationary regime points, and some may not. So, since the velocity vectors $\dot{z}(t)$ of the solutions $z(t)$ to the A problem belong to the boundary of the set $\bar{V}(z)$, the R problem approximates the SP problem if, during the motion, the vectors $\dot{z}(t)$ remain among the boundary quasi-stationary regime points.

Naturally, the approximation takes place if the set $\bar{V}(z)$ consists of the stationary regime points only.

THEOREM 4.3. *Under the conditions of Theorem 4.1, let the set $f(z, \bar{V}(z), U)$ be convex with any $z \in W$, where $f(z, y, u) \stackrel{\text{def}}{=} \{f_1(z, y, u), f_2(z, y, u)\}$. Then the R problem approximates the SP one.*

Proof. The proof follows from Theorem 4.2 and Proposition 3.2.

Example 4.3. Suppose that relations (4.9), (4.10) take place, and, moreover, the set U is convex and

$$(4.13) \quad f_1(z, y, u) = A_1(z)y + B_1(z)u + F_1(z).$$

Then (see Example 3.4) the conditions of Theorem 4.3 are satisfied, and the SP problem is approximated by the R problem. Note that similar results for the systems linear in fast variables and controls were obtained via the boundary layer method in Dmitriev [5], O'Malley [22], and Sannuti [26].

It is obvious that the statements of Corollary 2.1 are true under the conditions of Theorem 4.1. These statements concern the optimal control problems with the functional depending only on the slow variables. Let us consider which form is taken by results in the problems with functionals

$$(4.14) \quad \inf \Phi(z(1), y(1)).$$

The problem of minimization (4.14) on the set of the admissible trajectories of the singularly perturbed system (1.1), (1.2) is referred to as the SP-y problem to distinguish it from that of minimization (2.10). The optimal value of this problem is denoted by Φ_ϵ . We now show that the SP-y problem is approximated by the A problem with the functional in (2.10) defined as follows:

$$(4.15) \quad G(z) \stackrel{\text{def}}{=} \min\{\Phi(z, \eta) | \eta \in \bar{Y}(z)\}.$$

THEOREM 4.4. *Suppose that function $\Phi(z, y)$ is continuous and that the assumptions of Theorem 4.1 are true. Then $\Phi_\epsilon \rightarrow G_a$ as $\epsilon \rightarrow 0$, where G_a is the optimal value of the A problem with $G(z)$ defined as in (4.15). If $z(t)$ is a solution to such an A problem, an ϵ -suboptimal control for the SP-y problem is constructed according to the formula*

$$(4.16) \quad u_\epsilon(t) = \begin{cases} u_{z(\cdot)}(t), & t \in [0, 1 - \Delta_\epsilon], \\ u_\eta(\frac{t}{\epsilon}), & t \in [1 - \Delta_\epsilon, 1]. \end{cases}$$

Here $u_{z(\cdot)}(t)$ is z -approximating control; Δ_ϵ is an arbitrary function of ϵ such that

$$(4.17) \quad \lim_{\epsilon \rightarrow 0} \Delta_\epsilon = 0, \lim_{\epsilon \rightarrow 0} \epsilon^{-1} \Delta_\epsilon = \infty;$$

$u_\eta(\tau)$ is a control that when, using in the associated system (2.1), provides the validity of the inequality

$$(4.18) \quad \|y_{z(1)}(\epsilon^{-1}) - \eta(z(1))\| \leq 2C\xi(\epsilon^{-1}\Delta_\epsilon),$$

where C is defined in (3.7); $\eta(z)$ is a solution to the minimization problem in the right-hand side of (4.15); $y_{z(1)}(\tau)$ is the solution to the associated system (2.1) taken with $z = z(1)$ and considered on the interval $[\epsilon^{-1}(1 - \Delta_\epsilon), \epsilon^{-1}]$ with the control $u_\eta(\tau)$ and with the initial values $y_{z(1)}(\epsilon^{-1}(1 - \Delta_\epsilon)) = y_\epsilon(1 - \Delta_\epsilon)$; $y_\epsilon(t)$ are y -components of the solution $\{z_\epsilon(t), y_\epsilon(t)\}$ to system (1.1), (1.2) obtained with the control $u_{z(\cdot)}(t)$ (the existence of the control $u_\eta(\tau)$ follows from (4.3)).

Proof. The proof is given in §5.

Similar results for linear systems are obtained in Dmitriev [6], Dontchev [7], and Pervozvansky, and Gaitsgory [23].

Let us note in conclusion that statements of Theorems 4.1 and 4.3 permit us to interpret the fast variables as playing a role of some additional controls with respect to the slow ones. Statement (iii) of Theorem 4.1 may be considered to establish an asymptotic convexification of the slow variables velocities set. Being connected with the opportunity to rapidly use oscillating controls and fast variables, this statement is similar to those on convexification via sliding regimes or relaxed controls (see Gamkrelidze [13], Warga [30], Young [31], and Artstein [1] for another type of correlation between rapid oscillations and relaxed controls). Theorem 4.3 is close by its form to the well-known Filippov theorem [8] on the existence of the optimal control. It provides conditions “making unnecessary” the use of rapidly oscillating controls and fast variables to improve the value of the functional.

Theorem 4.4 is also connected with the special role of the fast variables. With the almost unchanged slow ones, they may reach a neighborhood of any point of the limit reachability set of the associated system and, in particular, approximate “the most profitable” point of this set defined as the solution to (4.15).

5. Proofs of Lemma 2.1 and Theorem 4.4. During the proof of Lemma 2.1, the following statement is used repeatedly.

PROPOSITION 5.1. *Let N_ϵ be a function of ϵ with its values being natural numbers tending to infinity as ϵ tends to zero. Then, if nonnegative numbers Δ_l satisfy the inequality $\Delta_{l+1} \leq \Delta_l + L_1 N_\epsilon^{-1} \Delta_l + \phi(\epsilon) N_\epsilon^{-1}$ with $l = 0, 1, \dots, k < N_\epsilon$, they also satisfy the inequality $\Delta_l \leq \phi(\epsilon) L_1^{-1} e^{L_1}$ with $l = 0, 1, \dots, k+1$, where $\Delta_0 \stackrel{\text{def}}{=} 0$, $L_1 > 0$, $\phi(\epsilon) \geq 0$.*

Proof. The proof follows from the fact that the numbers Δ_l satisfy the inequality $\Delta_l \leq \tilde{\Delta}_l$, $l = 0, 1, \dots, k+1$, where $\tilde{\Delta}_l$ are defined via the difference equation $\tilde{\Delta}_{l+1} = \tilde{\Delta}_l + L_1 N_\epsilon^{-1} \tilde{\Delta}_l + \phi(\epsilon) N_\epsilon^{-1}$, $\tilde{\Delta}_0 \stackrel{\text{def}}{=} 0$ and also from the estimate $\tilde{\Delta}_l \leq \phi(\epsilon) L_1^{-1} e^{L_1}$.

Proof of Lemma 2.1. For convenience, let us rewrite system (1.1), (1.2) in the stretched timescale $\tau = t\epsilon^{-1}$ as

$$(5.1) \quad \dot{z} = \epsilon f_1(z, y, u), \quad z(0) = z_0,$$

$$(5.2) \quad \dot{y} = f_2(z, y, u), \quad y(0) = y_0,$$

with $\tau \in [0, \epsilon^{-1}]$. Let $\{z_\epsilon(\tau), y_\epsilon(\tau)\}$ be the solution to system (5.1), (5.2) obtained with some admissible control $u(\tau)$. We should construct a solution $z(t)$ to the differential inclusion (2.5), satisfying the inequality

$$(5.3) \quad \max_{\tau \in [0, \epsilon^{-1}]} \|z_\epsilon(\tau) - z(\tau\epsilon)\| \leq \mu(\epsilon),$$

where $\mu(\epsilon)$ tends to zero as ϵ tends to zero.

Let us introduce the following notation:

$$S_\epsilon = \kappa \ln \epsilon^{-1}; \tau_l = l S_\epsilon, l = 0, 1, \dots, N_\epsilon \stackrel{\text{def}}{=} \lceil (\epsilon S_\epsilon)^{-1} \rceil; \tau_{N_\epsilon+1} \stackrel{\text{def}}{=} \epsilon^{-1},$$

where a positive number κ is specified below. Compare vectors $\{z_\epsilon(\tau_l)\}$ with the vectors z_l defined as the solution to the difference equation

$$(5.4) \quad z_{l+1} = z_l + \epsilon \int_{\tau_l}^{\tau_{l+1}} f_1(z_l, y_{z_l}(\tau), u(\tau)) d\tau,$$

where $y_{z_l}(\tau)$ is the solution to the associated system (2.1) obtained on the interval $[\tau_l, \tau_{l+1}]$ with the control $u(\tau)$ and the initial values $y_{z_l}(\tau_l) = y_\epsilon(\tau_l)$, and with $z = z_l$. Denote by $y_{z_\epsilon(\tau_l)}(\tau)$ the solution to the same system obtained with the same control and with the same initial values, but with $z = z_\epsilon(\tau_l)$. Subtracting (5.4) from the relation

$$(5.5) \quad z_\epsilon(\tau_{l+1}) = z_\epsilon(\tau_l) + \epsilon \int_{\tau_l}^{\tau_{l+1}} f_1(z_\epsilon(\tau), y_\epsilon(\tau), u(\tau)) d\tau,$$

we may write

$$(5.6) \quad \begin{aligned} \Delta_{l+1} \leq \Delta_l + \epsilon \int_{\tau_l}^{\tau_{l+1}} \|f_1(z_\epsilon(\tau), y_\epsilon(\tau), u(\tau)) - f_1(z_\epsilon(\tau_l), y_{z_\epsilon(\tau_l)}(\tau), u(\tau))\| d\tau \\ + \epsilon \left\| \int_{\tau_l}^{\tau_{l+1}} (f_1(z_\epsilon(\tau_l), y_{z_\epsilon(\tau_l)}(\tau), u(\tau)) - f_1(z_l, y_{z_l}(\tau), u(\tau))) d\tau \right\|, \end{aligned}$$

where $\Delta_l \stackrel{\text{def}}{=} \|z_\epsilon(\tau_l) - z_l\|$, $\Delta_0 \stackrel{\text{def}}{=} 0$. From Assumption 2.5, it follows that

$$(5.7) \quad \epsilon \left\| \int_{\tau_l}^{\tau_{l+1}} (f_1(z_\epsilon(\tau_l), y_{z_\epsilon(\tau_l)}(\tau), u(\tau)) - f_1(z_l, y_{z_l}(\tau), u(\tau))) d\tau \right\| \leq C \Delta_l \epsilon S_\epsilon.$$

By virtue of Assumption 2.3,

$$(5.8) \quad \|z_\epsilon(\tau) - z_\epsilon(\tau_l)\| \leq \epsilon S_\epsilon M, \quad \forall \tau \in [\tau_l, \tau_{l+1}],$$

$$\begin{aligned} \|y(\tau) - y_{z_\epsilon(\tau_l)}(\tau)\| &\leq \int_{\tau_l}^{\tau} \|f_2(z_\epsilon(s), y_\epsilon(s), u(s)) - f_2(z_\epsilon(\tau_l), y_{z_\epsilon(\tau_l)}(s), u(s))\| ds \\ &\leq L \int_{\tau_l}^{\tau} (\|z_\epsilon(s) - z_\epsilon(\tau_l)\| + \|y_\epsilon(s) - y_{z_\epsilon(\tau_l)}(s)\|) ds \\ &\leq LM\epsilon S_\epsilon^2 + L \int_{\tau_l}^{\tau} \|y_\epsilon(s) - y_{z_\epsilon(\tau_l)}(s)\| ds. \end{aligned}$$

Using the Gronwall–Bellman lemma, we obtain from here that

$$\begin{aligned} \|y(\tau) - y_{z_\epsilon(\tau_l)}(\tau)\| &\leq LM\epsilon S_\epsilon^2 e^{LS_\epsilon} \quad \forall \tau \in [\tau_l, \tau_{l+1}] \Rightarrow \\ &\epsilon \int_{\tau_l}^{\tau_{l+1}} \|f_1(z_\epsilon(\tau), y_\epsilon(\tau), u(\tau)) - f_1(z_\epsilon(\tau_l), y_{z_\epsilon(\tau_l)}(\tau), u(\tau))\| d\tau \\ &\leq L(\epsilon S_\epsilon)(\epsilon S_\epsilon M + LM\epsilon S_\epsilon^2 e^{LS_\epsilon}). \end{aligned}$$

We suppose in what follows that the coefficient κ in the definition of the function S_ϵ satisfies the inequality $\kappa < L^{-1}$. Then, as may be easily verified, the right-hand side of the last inequality is majorized with sufficiently small ϵ by the function $(\epsilon S_\epsilon)\epsilon^{(1-L\kappa)/2}$, which, together with (5.7), permits us to rewrite (5.6) in the following form: $\Delta_{l+1} \leq \Delta_l + CN_\epsilon^{-1}\Delta_l + \epsilon^{(1-L\kappa)/2}N_\epsilon^{-1}$, where it is also considered that $\epsilon S_\epsilon \leq N_\epsilon^{-1}$. Using this inequality and Proposition 5.1, we may obtain that

$$(5.9) \quad \Delta_l \stackrel{\text{def}}{=} \|z_\epsilon(\tau_l) - z_l\| \leq \epsilon^{(1-L\kappa)/2} C^{-1} e^C \quad \forall l = 0, 1, \dots, N_\epsilon.$$

In accordance with Assumption 2.1, there exist vectors $v_l \in \bar{V}(z_l)$, $l = 0, 1, \dots, N_\epsilon - 1$ such that the estimates $\|S_\epsilon^{-1} \int_{\tau_l}^{\tau_{l+1}} f_1(z_l, y_{z_l}(\tau), u(\tau)) d\tau - v_l\| \leq \gamma(S_\epsilon)$ are true if ϵ is small enough. Define the sequence of the vectors ζ_l as the solution to the equation

$$(5.10) \quad \zeta_{l+1} = \zeta_l + \epsilon S_\epsilon v_l, \quad l = 0, 1, \dots, N_\epsilon - 1; \quad \zeta_0 \stackrel{\text{def}}{=} 0.$$

After subtracting it from (5.4), we obtain that

$$(5.11) \quad \begin{aligned} \|z_{l+1} - \zeta_{l+1}\| &\leq \|z_l - \zeta_l\| + \epsilon S_\epsilon \gamma(S_\epsilon) \leq \|z_l - \zeta_l\| + N_\epsilon^{-1} \gamma(S_\epsilon) \Rightarrow \\ \|z_l - \zeta_l\| &\leq \gamma(S_\epsilon) \quad \forall l = 0, 1, \dots, N_\epsilon. \end{aligned}$$

Define the piecewise linear function $\zeta(t)$ according to the formula

$$\zeta(t) = \begin{cases} \zeta_l + (t - t_l)v_l, & t \in [t_l, t_{l+1}], l = 0, 1, \dots, N_\epsilon - 2, \\ \zeta_{N_\epsilon-1} + (t - t_{N_\epsilon-1})v_{N_\epsilon-1}, & t \in [t_{N_\epsilon-1}, 1], \end{cases}$$

where $t_l \stackrel{\text{def}}{=} \epsilon \tau_l$. Let us estimate the value $\rho(\dot{\zeta}(t), \bar{V}(\zeta(t)))$. Note first that from (2.8) it follows that

$$(5.12) \quad \max\{\|\eta\| \mid \eta \in \bar{V}(z)\} \leq M \quad \forall z \in W,$$

and from (2.9) it can be derived that

$$(5.13) \quad \rho(\bar{V}(z^1), \bar{V}(z^2)) \leq C\|z^1 - z^2\| \quad \forall z^1, z^2 \in W.$$

These relations provide the validity of the inequalities

$$\begin{aligned} \text{dist}(\dot{\zeta}(t), \bar{V}(\zeta(t))) &= \text{dist}(v_l, \bar{V}(\zeta(t))) \leq \text{dist}(v_l, \bar{V}(z_l)) + \rho(\bar{V}(z_l), \bar{V}(\zeta_l)) \\ &\quad + \rho(\bar{V}(\zeta_l), \bar{V}(\zeta(t))) \leq C(\gamma(S_\epsilon) + \epsilon S_\epsilon M) \\ &\quad \forall t \in (t_l, t_{l+1}), l = 0, 1, \dots, N_\epsilon - 2; \text{dist}(\dot{\zeta}(t), \bar{V}(\zeta(t))) \\ &\leq C(\gamma(S_\epsilon) + 2\epsilon S_\epsilon M) \quad \forall t \in (t_{N_\epsilon-1}, 1), \end{aligned}$$

where $\rho(\cdot, \cdot)$ and $\text{dist}(\cdot, \cdot)$ are defined as follows:

$$\rho(V_1, V_2) \stackrel{\text{def}}{=} \max\left\{\sup_{\eta \in V_1} \text{dist}(\eta, V_2), \sup_{\eta \in V_2} \text{dist}(\eta, V_1)\right\}, \text{dist}(\eta, V_i) \stackrel{\text{def}}{=} \inf_{\eta' \in V_i} \|\eta - \eta'\|.$$

From the obtained inequalities and Filippov's result [9, Thm. 1], it follows that there exists a solution $z(t)$ to the differential inclusion (2.5) such that

$$\max_{t \in [0, 1]} \|z(t) - \zeta(t)\| \leq e^C(\gamma(S_\epsilon) + 2\epsilon S_\epsilon M).$$

This estimate, together with (5.11), (5.9), and (5.8), allows us to verify that the indicated solution $z(t)$ satisfies inequality (5.3) with $\mu(\epsilon) = O(\gamma(S_\epsilon)) + O(\epsilon^{(1-L\kappa)/2})$.

Now let $z(t)$ be an arbitrary solution to the differential inclusion (2.5), and let us construct z -approximating control $u(\tau)$. By virtue of (5.12), (5.13) we may write with $t \in [t_l, t_{l+1}]$, $l = 0, 1, \dots, N_\epsilon - 1$

$$\dot{z}(t) \in \bar{V}(z(t)) \subset \bar{V}(z(t_l)) + C\|z(t) - z(t_l)\| \bar{B} \subset \bar{V}(z(t_l)) + CM(\epsilon S_\epsilon) \bar{B},$$

where \bar{B} is the closed ball in R^{n_1} with the center in the origin and with the unit radius. Consequently,

$$\begin{aligned} (5.14) \quad (\epsilon S_\epsilon)^{-1} \int_{t_l}^{t_{l+1}} \dot{z}(t) dt &\in \bar{V}(z(t_l)) + CM(\epsilon S_\epsilon) \bar{B} \\ \Rightarrow \text{dist}((\epsilon S_\epsilon)^{-1} \int_{t_l}^{t_{l+1}} \dot{z}(t) dt, \bar{V}(z(t_l))) &\leq CM(\epsilon S_\epsilon). \end{aligned}$$

Define the vectors v_l , $l = 0, 1, \dots, N_\epsilon - 1$ as the projections of the vectors

$$(\epsilon S_\epsilon)^{-1} \int_{t_l}^{t_{l+1}} \dot{z}(t) dt$$

onto the sets $\bar{V}(z(t_l))$: $v_l \stackrel{\text{def}}{=} \text{argmin}\{\|(\epsilon S_\epsilon)^{-1} \int_{t_l}^{t_{l+1}} \dot{z}(t) dt - v\| \mid v \in \bar{V}(z(t_l))\}$. Suppose that the control $u(\tau)$ is defined on the interval $[0, \tau_l]$, $l < N_\epsilon$ and extend its definition to the interval $(\tau_l, \tau_{l+1}]$. Denote by $\{z_\epsilon(\tau), y_\epsilon(\tau)\}$ the trajectory of system (5.1), (5.2)

obtained with the use of the control $u(\tau)$ on the interval $[0, \tau_l]$. Define the control $u(\tau)$ on the interval $(\tau_l, \tau_{l+1}]$ as that providing the fulfillment of the inequality

$$(5.15) \quad \|S_\epsilon^{-1} \int_{\tau_l}^{\tau_{l+1}} f_1(z(t_l), y_{z(t_l)}(\tau), u(\tau)) d\tau - v_l\| \leq 2\gamma(S_\epsilon),$$

where $y_{z(t_l)}(\tau)$ is the solution to the associated system obtained with the control $u(\tau)$ and the initial values $y_{z(t_l)}(\tau_l) = y_\epsilon(\tau_l)$ and with $z = z(t_l)$. The existence of such control follows from Assumption 2.1. Thus the control $u(\tau)$ can be defined on the interval $[0, \tau_{N_\epsilon}]$. On the interval $[\tau_{N_\epsilon}, 1]$, we complete the definition in an arbitrary way. Let us show that the trajectory $\{z_\epsilon(t), y_\epsilon(t)\}$ of system (5.1), (5.2) obtained with this control satisfies (5.3). Define the sequence $\{\zeta_l\}$ as the solution to the difference equation (5.10) and subtract this equation from the relation $z(t_{l+1}) = z(t_l) + \int_{t_l}^{t_{l+1}} \dot{z}(t) dt$. By virtue of (5.14), we obtain that

$$(5.16) \quad \begin{aligned} \|z(t_{l+1}) - \zeta_{l+1}\| &\leq \|z(t_l) - \zeta_l\| + \epsilon S_\epsilon \|(\epsilon S_\epsilon)^{-1} \int_{t_l}^{t_{l+1}} \dot{z}(t) dt - v_l\| \leq \|z(t_l) - \zeta_l\| + CM(\epsilon S_\epsilon)^2 \\ &\rightarrow \|z(t_l) - \zeta_l\| \leq CM\epsilon S_\epsilon \quad \forall l = 0, 1, \dots, N_\epsilon. \end{aligned}$$

Define the sequence of the vectors $\{z_l\}$ in accordance with (5.4). Subtracting (5.10) from (5.4), we have that

$$\begin{aligned} \|z_{l+1} - \zeta_{l+1}\| &\leq \|z_l - \zeta_l\| + \epsilon S_\epsilon \|S_\epsilon^{-1} \int_{\tau_l}^{\tau_{l+1}} f_1(z_l, y_{z_l}(\tau), u(\tau)) d\tau \\ &\quad - S_\epsilon^{-1} \int_{\tau_l}^{\tau_{l+1}} f_1(z(t_l), y_{z(t_l)}(\tau), u(\tau)) d\tau\| + \epsilon S_\epsilon \|S_\epsilon^{-1} \\ &\quad \cdot \int_{\tau_l}^{\tau_{l+1}} f_1(z(t_l), y_{z(t_l)}(\tau), u(\tau)) d\tau - v_l\|, \end{aligned}$$

where $y_{z(t_l)}(\tau)$ is the solution to the associated system (2.1) obtained with the control $u(\tau)$, with the initial values $y_{z(t_l)}(\tau_l) = y_\epsilon(\tau_l)$ and with $z = z(t_l)$. From Assumption 2.5 and (5.16), it follows that

$$\begin{aligned} S_\epsilon^{-1} \left\| \int_{\tau_l}^{\tau_{l+1}} f(z_l, y_{z_l}(\tau), u(\tau)) d\tau - \int_{\tau_l}^{\tau_{l+1}} f_1(z(t_l), y_{z(t_l)}(\tau), u(\tau)) d\tau \right\| &\leq C \|z_l - z(t_l)\| \\ &\leq C \|z_l - \zeta_l\| + C^2 M \epsilon S_\epsilon. \end{aligned}$$

Using this and (5.15), we may write

$$\|z_{l+1} - \zeta_{l+1}\| \leq \|z_l - \zeta_l\| + CN_\epsilon^{-1} \|z_l - \zeta_l\| + (C^2 M \epsilon S_\epsilon + 2\gamma(S_\epsilon)) N_\epsilon^{-1},$$

which, on the basis of Proposition 5.1, permits us to establish that

$$(5.17) \quad \|z_l - \zeta_l\| \leq C^{-1} (C^2 M \epsilon S_\epsilon + 2\gamma(S_\epsilon)) e^C \quad \forall l = 0, 1, \dots, N_\epsilon.$$

By just the same reasoning as above, we may verify the validity of (5.9), which, together with (5.17), (5.16), (5.12), and (2.8), allows us to obtain (5.3). \square

Note that the proof considered is in many respects similar to that given in Plotnikov [25] for a particular case where $n_2 = 1$, $f_2(z, y, u) \equiv 1$.

Proof of Theorem 4.4. Again, it is more convenient to deal with the system in the stretched timescale (5.1), (5.2), and we rewrite the functional (4.14) in the form

$$(5.18) \quad \inf \Phi(z(\epsilon^{-1}), y(\epsilon^{-1})).$$

Control (4.16) is rewritten in this timescale as

$$(5.19) \quad u_\epsilon(\tau) = \begin{cases} u_{z(\cdot)}(\tau), & \tau \in [0, \epsilon^{-1}(1 - \Delta_\epsilon)), \\ u_{\eta(\cdot)}(\tau), & \tau \in [\epsilon^{-1}(1 - \Delta_\epsilon), \epsilon^{-1}]. \end{cases}$$

Since the control $u_\epsilon(\tau)$ coincides with z -approximating one $u_{z(\cdot)}(\tau)$ on the interval $[0, \epsilon^{-1}(1 - \Delta_\epsilon))$, from (5.3), (2.8), and (5.12), it follows that

$$(5.20) \quad \max_{\tau \in [\epsilon^{-1}(1 - \Delta_\epsilon), \epsilon^{-1}]} \|z_\epsilon(\tau) - z(1)\| \leq \mu(\epsilon) + 2M\Delta_\epsilon,$$

where $\{z_\epsilon(\tau), y_\epsilon(\tau)\}$ is the trajectory of (5.1), (5.2) obtained with control (5.19). On the basis of Lemma 4.1 and (4.18), we may write

$$\begin{aligned} \|y_\epsilon(\epsilon^{-1}) - \eta(z(1))\| &\leq \|y_\epsilon(\epsilon^{-1}) - y_{z(1)}(\epsilon^{-1})\| + \|y_{z(1)}(\epsilon^{-1}) - \eta(z(1))\| \\ &\leq C_1(\mu(\epsilon) + 2M\Delta_\epsilon) + 2C\xi(\epsilon^{-1}\Delta_\epsilon). \end{aligned}$$

From this and from (5.20), it follows that the value of the functional (5.18) obtained with control (5.19) tends to G_a as ϵ tends to zero, as follows:

$$(5.21) \quad \lim_{\epsilon \rightarrow 0} \Phi(z_\epsilon(\epsilon^{-1}), y_\epsilon(\epsilon^{-1})) = \Phi(z(1), \eta(z(1))) \stackrel{\text{def}}{=} G_a.$$

Now let $\tilde{u}(\tau)$ be an arbitrary admissible control and $\{\tilde{z}_\epsilon(\tau), \tilde{y}_\epsilon(\tau)\}$ the corresponding trajectory of system (5.1), (5.2). According to Theorem 4.1, there exists a solution $\tilde{z}(t)$ to the differential inclusion (2.5), which satisfies the following inequality similar to (5.20):

$$(5.22) \quad \max_{\tau \in [\epsilon^{-1}(1 - \Delta_\epsilon), \epsilon^{-1}]} \|\tilde{z}_\epsilon(\tau) - \tilde{z}(1)\| \leq \mu(\epsilon) + 2M\Delta_\epsilon.$$

Denote by $y_{\tilde{z}(1)}(\tau)$ the solution to the associated system (2.1) obtained when using the control $\tilde{u}(\tau)$ on the interval $[\epsilon^{-1}(1 - \Delta_\epsilon), \epsilon^{-1}]$ with the initial values $y_{\tilde{z}(1)}(\epsilon^{-1}(1 - \Delta_\epsilon)) = \tilde{y}_\epsilon(\epsilon^{-1}(1 - \Delta_\epsilon))$ and with $z = \tilde{z}(1)$. By virtue of Lemma 4.1 and (5.22), $\|\tilde{y}_\epsilon(\epsilon^{-1}) - y_{\tilde{z}(1)}(\epsilon^{-1})\| \leq C_1(\mu(\epsilon) + 2M\Delta_\epsilon)$. On the other hand,

$$y_{\tilde{z}(1)}(\epsilon^{-1}) \in Y(\tilde{z}(1), \epsilon^{-1}\Delta_\epsilon, \tilde{y}_\epsilon(\epsilon^{-1}(1 - \Delta_\epsilon))),$$

and, in accordance with (4.3), there exists a vector $\tilde{\eta} \in \bar{Y}(\tilde{z}(1))$ such that

$$(5.23) \quad \|\tilde{\eta} - y_{\tilde{z}(1)}(\epsilon^{-1})\| \leq 2C\xi(\epsilon^{-1}\Delta_\epsilon) \Rightarrow \|\tilde{\eta} - \tilde{y}_\epsilon(\epsilon^{-1})\| \leq 2C\xi(\epsilon^{-1}\Delta_\epsilon) + C_1(\mu(\epsilon) + 2M\Delta_\epsilon).$$

Since $\Phi(\tilde{z}(1), \tilde{\eta}) \geq G_a$, from (5.22), (5.23), it follows that

$$\liminf_{\epsilon \rightarrow 0} \Phi(\tilde{z}_\epsilon(\epsilon^{-1}), \tilde{y}_\epsilon(\epsilon^{-1})) = \liminf_{\epsilon \rightarrow 0} \Phi(\tilde{z}(1), \tilde{\eta}) \geq G_a.$$

As $\{\tilde{z}_\epsilon(\tau), \tilde{y}_\epsilon(\tau)\}$ is an arbitrary admissible trajectory of system (5.1), (5.2), we obtain from here that $\lim_{\epsilon \rightarrow 0} \inf \Phi_\epsilon \geq G_a$, which, together with (5.21), completes the proof. \square

Acknowledgment. The author thanks Professor J. Warga for the useful discussion of the results.

REFERENCES

- [1] Z. ARTSTEIN, *Rapid oscillations, chattering systems and relaxed controls*, SIAM J. Control Optim., 27 (1979), pp. 940–948.
- [2] J. E. BAILEY AND F. J. M. HORN, *Comparison between two sufficient conditions for improvement of an optimal steady-state process by periodic operation*, J. Optim. Theory Appl., 7 (1971).
- [3] A. BENSOUSSAN, *Perturbation Methods in Optimal Control Problems*, John Wiley, New York, 1989.
- [4] A. M. CIRLIN, V. S. BALAKIREV, AND E. G. DUDNIKOV, *Variational Methods in Optimization of Control Objects*, Energia, Moscow, 1976. (In Russian.)
- [5] M. G. DMITRIEV, *Continuity of the solutions to non-linear Maier optimal control problem with respect to singular perturbations*, Differentsial'nye Uravneniye i Primeneniye, 2 (1971), pp. 21–29. (In Russian.)
- [6] ———, *Boundary layer in optimal control problems*, Prikl. Mat. Mekh., 42 (1978), pp. 228–232. (In Russian.)
- [7] A. L. DONTCHEV, *Perturbations, Approximations and Sensitivity Analysis of Optimal Control Systems*, Lecture Notes in Control and Inform. Sci., Springer-Verlag, Berlin, 1983.
- [8] A. F. FILIPPOV, *To some questions of the theory of optimal control*, Vestnik Moskov. Univ. Ser. I Mat. Mekh., 2 (1959), pp. 25–32. (In Russian.)
- [9] ———, *Classical solutions to differential equations with multi-valued right-hand side*, Vestnik Moskov. Univ. Ser. I Mat. Mekh., 3 (1967), pp. 16–26. (In Russian.)
- [10] ———, *Differential Equations with Discontinuous Right-Hand Sides*, Kluwer Academic Publishers, Dordrecht, the Netherlands, 1988.
- [11] V. G. GAITSGORY, *Use of the averaging method in control problems*, Differential Equations (translated from Russian), 22 (1986), pp. 1290–1299.
- [12] ———, *Application of the averaging method for construction of suboptimal solutions to singularly perturbed optimal control problems*, Automat. Remote Control (translated from Russian), 46 (1986), pp. 1081–1088.
- [13] R. V. GAMKRELIDZE, *About sliding optimal regimes*, Soviet Math. Dokl., 143 (1962), pp. 1243–1245. (In Russian.)
- [14] A. H. GELIG, G. A. LEONOV, AND V. A. IAKUBOVICH, *Stability of Non-Linear Systems with Nonunique State of Equilibrium*, Nauka, Moscow, 1978. (In Russian.)
- [15] E. G. GILBERT, *Optimal periodic control: A general theory of necessary conditions*, SIAM J. Control Optim., 5 (1977), pp. 717–746.
- [16] G. GUARDABASSI, A. LOCATELLI, AND S. RINALDI, *Status of periodic optimization of dynamical systems*, J. Optim. Theory Appl., 14 (1974), pp. 11–20.
- [17] D. H. HADWIGER, *Vorlesungen uber Inhalt, Oberflache und Isoperimetrie*, Springer-Verlag, Berlin, 1957.
- [18] P. V. KOKOTOVIC, *Applications of singular perturbations techniques to control problems*, SIAM Rev., 26 (1984), pp. 501–550.
- [19] P. V. KOKOTOVIC, H. KHALIL, AND J. O'REILLY, *Singular Perturbations in Control Analysis and Design*, Academic Press, New York, 1986.
- [20] P. V. KOKOTOVIC, R. E. O'MALLEY, AND P. SANNUTI, *Singular perturbations and order reduction in control theory*, Automatica, 12 (1976), pp. 123–132.
- [21] R. E. O'MALLEY, *Introduction to Singular Perturbations*, Academic Press, New York, 1974.
- [22] ———, *Boundary layer methods for certain non-linear singularly perturbed optimal control problems*, J. Math. Anal. Appl., 45 (1974), pp. 468–484.
- [23] A. A. PERVOZVANSKY AND V. G. GAITSGORY, *Theory of Suboptimal Decisions*, Kluwer Academic Publishers, Dordrecht, the Netherlands, 1988.
- [24] V. A. PLOTNIKOV, *Averaging method for differential inclusions and its application to optimal control problems*, Differential Equations (translated from Russian), 15 (1979), pp. 1013–1018.
- [25] 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. 273–294.
- [26] P. SANNUTI, *Asymptotic solution to singularly perturbed optimal control problems*, Automatica, 10 (1974), pp. 183–194.
- [27] A. N. TICHONOV, *Systems of differential equations containing small parameters near deriva-*

- tives, Mat. Sb., 31 (1952), pp. 575–586. (In Russian.)
- [28] A. B. VASIL'eva AND V. F. BUTUZOV, *Asymptotic Expansions of Solutions to Singularly Perturbed Equations*, Nauka, Moscow, 1973. (In Russian.)
- [29] V. M. VOLOSOV, *Averaging in systems of ordinary differential equations*, Uspekhi Mat. Nauk, 17 (1962), pp. 3–126. (In Russian.)
- [30] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
- [31] L. C. YOUNG, *Generalized curves and the existence of an attained absolute minimum in the calculus of variations*, C. R. Acad. Sci. Lett. Varsovie CIII, 30 (1937), pp. 212–234.

ON THE TIME-VARYING RICCATI DIFFERENCE EQUATION OF OPTIMAL FILTERING*

GIUSEPPE DE NICOLAO†

Abstract. This paper studies the time-varying Riccati difference equation (RDE) for the filtering problem. In particular, existence, stabilizability, and attractiveness properties of the real symmetric solutions that remain bounded on $(-\infty, +\infty)$ (infinite-time solutions) are investigated. Under the assumption of uniform detectability, conditions for the existence of the maximal and stabilizing solutions are given. Analogous results are worked out for the minimal and antistabilizing solutions by making reference to the uniform antidetectability notion. Moreover, it is shown that, under uniform observability, the set of all symmetric infinite-time solutions constitute an infinite number of lattices with common minimal and maximal elements.

Key words. Riccati difference equation, linear time-varying systems, Kalman filtering, optimal control, stabilizability and detectability

AMS(MOS) subject classifications. 49N10, 93C50, 93C55, 93E11, 93E20

1. Introduction. Since the early 1960s, it has become clear that the matrix Riccati equation is the keystone of a number of filtering and control problems. Thirty years later, due to the effort of a multitude of authors, the theory of the time-invariant Riccati equation, although yet in progress, appears rich and consistent. It would be beyond the scope of the present Introduction to mention all the relevant contributions, and the following summary does not advance any claim of completeness. To make life easy, we will only refer to the Riccati equation for the filtering problem, taking for granted that the results extend by duality to the optimal control Riccati equation.

In the pioneering works of Kalman [1] and Bucy [2], the algebraic Riccati equation (ARE) was investigated under the assumptions of controllability and observability to derive the uniqueness of the symmetric positive semidefinite (SPS) solution, the stability of the closed loop system, and the asymptotic convergence properties. The relaxation of these hypotheses to the weaker ones of stabilizability and detectability is due to the works of Wonham [3] and Kucera [4], in continuous time, and Caines and Mayne [5] in discrete time. Starting from the early 1970s, there was a growing interest in the study of the nonstabilizable case. We can mention the contributions of J. C. Willems [6], Martensson [7], Kucera [8], Molinari [9], and, more recently, Callier and J. L. Willems [10], Chan, Goodwin, and Sin [11], and De Souza, Gevers, and Goodwin [12]. In the nonstabilizable case, the ARE admits more than one SPS solution, and the problem of the classification of the solutions arises. It has been shown that under detectability assumptions, a maximal solution exists, and conditions for this solution to be stabilizing have been provided. The convergence of the SPS time-varying solutions of the time-invariant differential or difference Riccati equation toward the SPS solutions of the ARE was also studied. The classification of all the real symmetric solutions (positive, negative, or even nondefinite) of the ARE calls for the works of J. C. Willems [6], Coppel [13], Callier and J. L. Willems [10], and Shayman [14]. A major result was the characterization of all the real symmetric solutions as a distributive lattice having

* Received by the editors May 9, 1990; accepted for publication (in revised form) April 12, 1991. Research for this paper was supported by the Centro di Teoria dei Sistemi of the Italian National Research Council (CNR) and by the Ministry of University and Scientific and Technological Research (MURST)—Project “Model Identification, Systems Control, Signal Processing.”

† Centro Teoria dei Sistemi—Consiglio Nazionale delle Ricerche, c/o Dipartimento di Elettronica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy.

the maximal and minimal solution as extremal elements. When we turn to the time-varying Riccati equation, much of the richness of the time-invariant case is lost. By suitably defining the notions of uniform reachability and observability, Kalman [1] and Bucy [2] (Deyst and Price [15] for the discrete-time case) showed that their main results also hold true in the time-varying case. Indeed, they proved that, under these assumptions, there is a unique SPS “moving equilibrium” and that the closed-loop system is asymptotically stable for any solution, independently of the (SPS) initial condition. An attempt to extend the RDE analysis with stabilizability and detectability to the time-varying discrete-time case was made by Hager and Horowitz [16] and Anderson and Moore [17]. The major contributions in [17] were (a) the exploration of equivalent definitions of uniform stabilizability and detectability; (b) the demonstration of a time-varying version of the Lyapunov lemma under stabilizability assumptions; and (c) the exponential stability of the closed-loop system under stabilizability and detectability. Only recently, in the context of continuous-time infinite-dimensional systems, Da Prato and Ichikawa [18] have established some existence and convergence results for the SPS solutions of the time-varying Riccati equation. In particular, a necessary and sufficient condition for the existence of a bounded SPS solution has been proved together with some results relative to the stabilizing and maximal solutions. However, to the author’s knowledge, the study of the negative semidefinite solutions, together with the notions of minimal and antistabilizing solution and the classification of all the symmetric solutions, has remained an unexplored region, at least for the most general time-varying case. Only for the class of periodically time-varying systems, the theory of the periodic Riccati equation appears almost as complete as its stationary counterpart; see, e.g., [19]–[22].

The purpose of this paper is to fill some of the gaps between the theory of the stationary RDE and the theory of the time-varying RDE. The analysis will be carried out only in discrete time, but a derivation of the analogous continuous time results should, in principle, be possible. The attention will be focused on the solutions remaining bounded on $(-\infty, +\infty)$, which will be termed “infinite-time” solutions. Conversely, the solutions that are obtained starting with a given initial condition will be called “finite-time solutions.” The symmetric infinite-time solutions will be shown to enjoy many of the properties that in the stationary case characterize the constant solutions of the RDE. In particular, uniform detectability and stabilizability guarantee the existence of a unique SPS infinite-time solution, which is attractive for all the SPS finite-time solutions. When the stabilizability assumption is removed, the uniqueness falls. However, it can be proved that a maximal infinite-time solution exists and is attractive for a certain set of finite-time solutions. By means of a device that allows us to reduce the study of the symmetric negative semidefinite (SNS) solutions of the RDE to the study of the SPS solutions of a suitably modified RDE, the SNS solutions are explored, also. In this context, the notions of antistabilizability, antidetectability, SNS, antistabilizing, and minimal solution take the place of stabilizability, detectability, SPS, stabilizing, and maximal solution. Finally, a classification of all the symmetric solutions is provided. It is shown that, under certain assumptions, the set of all symmetric infinite-time solutions can be grouped in an infinite number of isomorphic distributive lattice that share the same maximal and minimal elements (the maximal and the minimal solution, respectively).

The layout of the paper is as follows. In § 2 some preliminary definitions are given and the notion of “infinite-time” solution is introduced. Sections 3 and 4 are devoted to the SPS and SNS solutions, respectively. In § 5 the results of the previous sections are put together to give a comprehensive picture of all the symmetric solutions and

the lattice structure is proved. Finally, the Appendix includes a number of auxiliary lemmas that are employed throughout the paper.

2. Preliminaries.

2.1. Structural properties. Consider the following discrete-time, time-varying, linear system:

$$(1a) \quad x(t+1) = A(t)x(t) + v(t),$$

$$(1b) \quad y(t) = C(t)x(t) + w(t),$$

where $A(t): Z \rightarrow R^{n \times n}$, $B(t): Z \rightarrow R^{n \times m}$, $C(t): Z \rightarrow R^{p \times n}$ are bounded matrices on $(-\infty, +\infty)$, and $v(\cdot)$ and $w(\cdot)$ are independent zero-mean white noises having bounded covariance matrices $Q(\cdot)$ and $R(\cdot)$, respectively, with $Q(t) \geq 0$ and $R(t) > 0$, for all t . Moreover, it will also be assumed that $R(\cdot)^{-1}$ is bounded. The transition matrix of $A(t)$ will be denoted by $\Phi(t_2, t_1)$, $t_2 \geq t_1$.

A first important notion regarding system (1) is provided in the following definition.

DEFINITION 1 (exponential stability). $A(\cdot)$ is said to be exponentially stable on $[t_0, t_f]$ if there exist positive constants α and β such that

$$\|\Phi(t_2, t_1)\| \leq \alpha e^{-\beta(t_2 - t_1)}, \quad t_0 \leq t_1 \leq t_2 \leq t_f.$$

$A(\cdot)$ is said to be exponentially stable if it is exponentially stable on $(-\infty, +\infty)$.

The above definition looks unusual. However, it will prove useful later in discussing the exact statement of the Lyapunov lemma.

It is well known that the notion of stability plays a key role in the analysis of SPS solutions of the RDE. It will prove that, when analyzing the SNS solutions, an analogous role is played by the antistability notion.

DEFINITION 2 (exponential antistability). $A(\cdot)$ is said to be exponentially antistable on $[t_0, t_f]$ if $A(t)$ is nonsingular for $t \in [t_0, t_f]$, and $A(\cdot)^{-1}$ is exponentially stable on $[t_0, t_f]$.

$A(\cdot)$ is said to be exponentially antistable if it is exponentially antistable on $(-\infty, +\infty)$.

In the study of filtering and control problems, suitable modifications of the concepts of reachability and observability, under the names of uniform reachability and uniform observability, have proved particularly effective; see, e.g., [1]. For stationary systems, stabilizability and detectability can be seen as the natural relaxations of the controllability and observability notions. To extend to time-varying systems some results already known for the time-invariant case, the notions of uniform stabilizability and detectability were introduced in [16], [17]. In particular, in the latter reference, an exploration of equivalent definitions, together with an analysis of the duality between stabilizability and detectability, was carried out. In the following, we will use the uniform stabilizability criterion reported below. The definition of uniform antistabilizability is also introduced.

Uniform stabilizability and detectability criterion [17]. The pair $(A(\cdot), B(\cdot))$ $[(A(\cdot), C(\cdot))]$ is uniformly stabilizable (detectable) if and only if there exists a bounded matrix function $K(\cdot)$ such that $A(\cdot) + B(\cdot)K(\cdot)$ $[A(\cdot) + K(\cdot)C(\cdot)]$ is exponentially stable.

DEFINITION 3 (uniform antistabilizability and antidetectability). The pair $(A(\cdot), B(\cdot))$ $[(A(\cdot), C(\cdot))]$ is said to be uniformly antistabilizable (antidetectable) if

there exists a bounded matrix function $K(\cdot)$ such that $A(\cdot) + B(\cdot)K(\cdot) [A(\cdot) + K(\cdot)C(\cdot)]$ is exponentially antistable.

It can be shown that uniform reachability (observability) implies both uniform stabilizability (detectability) and uniform antistabilizability (antidetectability). In other words, if the pair $(A(\cdot), B(\cdot))$ is uniformly reachable, it is simultaneously uniformly stabilizable and antistabilizable.

Remark 1. Recently, in [23], a counterexample was adduced that seemed to infringe the uniform stabilizability criterion. Precisely, it was claimed that uniform stabilizability, as defined in [17], does not imply the existence of an exponentially stabilizing gain. In point of fact, as shown in [24], the time-varying pair $(A(\cdot), C(\cdot))$ considered in the counterexample, in contrast to the assertion of [23], does not satisfy the uniform detectability definition given in [17], so that the purported confutation is not valid.

2.2. Lyapunov and Riccati equations. It is well known that the state covariance matrix $\text{Var}[x(t)]$ obeys the following Lyapunov difference equation (LDE):

$$(2) \quad X(t+1) = A(t)X(t)A(t)' + Q(t).$$

Precisely, assume that $x(t_0)$ is a random variable independent of $v(t)$ and $w(t)$, $t \geq t_0$, and let $X(t_0) = \text{Var } x(t_0)$. Then, $X(t) = \text{Var}[x(t)]$, $t \geq t_0$.

The problem of finding the optimal (in the mean square sense) one-step prediction $x(t+1|t)$ is solved by means of the optimal Kalman predictor

$$x(t+1|t) = F(t)x(t|t-1) + K(t)y(t), \quad F(t) = A(t) - K(t)C(t),$$

$$K(t) = A(t)X(t)C(t)'[C(t)X(t)C(t)' + R(t)]^{-1},$$

where $X(t)$ is the solution of the RDE

$$(3) \quad \begin{aligned} X(t+1) = & A(t)X(t)A(t)' + Q(t) \\ & - A(t)X(t)C(t)'[C(t)X(t)C(t)' + R(t)]^{-1}C(t)X(t)A(t)', \end{aligned}$$

with initial condition $X(t_0) = \text{Var}[x(t_0)]$. The matrix function $F(\cdot)$ is the so-called closed-loop matrix, whereas $K(\cdot)$ is the celebrated Kalman gain. For each $t \geq t_0$, $X(t)$ provides the variance of the state prediction error.

In the stationary case, a vast amount of literature has flourished around the constant solutions of the RDE, which satisfy the so-called algebraic Riccati equation (ARE). Indeed, whenever such constant solutions are attractors for a class of nonconstant solutions, they provide an appealing tool for designing suboptimal constant predictors. Moreover, significant links between the constant solutions of the RDE and problems like spectral factorization and stochastic realization have been pointed out. It would seem that in the time-varying case there are no "privileged" solutions of the RDE, since all solutions are time-varying. However, for the particular class of periodically time-varying systems, it has been demonstrated that the periodic solutions of the periodic RDE play the same role as the constant solutions of the constant RDE [19]–[22]. In this paper, it will be shown that the analogy can be extended to the most general case of arbitrarily varying systems. Precisely, the time-varying equivalents of the constant solutions are provided by the solutions of the RDE that remain bounded on $(-\infty, +\infty)$. This leads to the following definitions, where the expressions "finite time" and "infinite time" refer to the initial condition.

DEFINITION 4 (finite-time solution of the RDE and LDE). By finite-time solution $X(\cdot)$ of the RDE [LDE], we mean a bounded matrix function $X(\cdot)$ defined over $[t_0, +\infty)$, which satisfies the RDE [LDE].

DEFINITION 5 (reversed finite-time solution of the RDE and LDE). By reversed finite-time solution $X(\cdot)$ of the RDE [LDE], we mean a bounded matrix function $X(\cdot)$ defined over $(-\infty, t_0]$, which satisfies the RDE [LDE].

DEFINITION 6 (infinite-time solution of the RDE and LDE). By infinite-time solution $X(\cdot)$ of the RDE [LDE], we mean a bounded matrix function $X(\cdot)$ defined over $(-\infty, +\infty)$, which satisfies the RDE [LDE].

In the following, we will mainly deal with infinite-time solutions. Therefore, whenever not explicitly stated, “solution of the RDE” will mean “infinite-time solution of the RDE.”

As already mentioned, some solutions of the ARE may enjoy the property of being attractive for a set of time-varying solutions. For example, in the stationary case, it is well known that, under stabilizability and detectability assumptions, all the SPS solutions of the RDE asymptotically converge to the unique SPS solution of the ARE. An extension of this notion to the time-varying case traces back to [1], where the term “moving equilibrium” was coined to denote the attractive time-varying SPS solution that exists under uniform reachability and observability. In our investigation of the convergence properties of the solutions of the time-varying RDE, the following generalization of the notion of moving equilibrium will prove useful.

DEFINITION 7 (moving equilibrium).

- (i) An infinite-time solution $\bar{X}(\cdot)$ of the RDE [LDE] is said to be a moving equilibrium for a certain set S of finite-time solutions of the RDE [LDE] if, for any finite-time solution $X(\cdot) \in S$, $\lim_{t \rightarrow \infty} X(t) - \bar{X}(t) = 0$.
- (ii) An infinite-time solution $\bar{X}(\cdot)$ of the RDE [LDE] is said to be a moving equilibrium for a certain set S of reversed finite-time solutions of the RDE [LDE] if, for any reversed finite-time solution $X(\cdot) \in S$, $\lim_{t \rightarrow -\infty} X(t) - \bar{X}(t) = 0$.

2.3. Extended Lyapunov lemma. In the stationary case, it is well known that there is a strict relationship between the stability properties of system (1a) and the constant SPS solutions of the LDE. Such a relationship is clarified by the renowned Lyapunov lemma, which, in its more general formulation, holds under the stabilizability assumption. The extension of the Lyapunov lemma to the time-varying case is provided in [25], under uniform reachability, and [17], under uniform stabilizability. By making reference to the notion of infinite-time solution, such a lemma can be given the following formulation.

LEMMA 1 (extended Lyapunov lemma). *Suppose that $(A(\cdot), Q(\cdot))$ is uniformly stabilizable and $A(\cdot)$ and $Q(\cdot)$ are bounded. Then, if there exists an infinite-time SPS solution $X(\cdot)$ of the LDE (2), $A(\cdot)$ is exponentially stable. Conversely, if $A(\cdot)$ is exponentially stable, there exists a unique infinite-time SPS solution $X(\cdot)$ of the LDE.*

Remark 2. Note that in [17] the Lyapunov lemma is correctly stated for the dual of (2), whereas the statement relative to (2) is imprecise [17, Thm. 4.3]. Indeed, it says (correctly) that, under uniform stabilizability, the existence of a finite-time SPS solution of the LDE on $[t_0, +\infty)$ implies the exponential stability of $A(\cdot)$ on $[t_0, +\infty)$. However, it is not true that the exponential stability of $A(\cdot)$ on $[t_0, +\infty)$ entails the existence of a *unique* finite-time SPS solution of the LDE on $[t_0, +\infty)$. A counterexample can be easily found by taking $A(t) = 0.5$, $Q(t) = 1$, for all t , and two different SPS initial conditions for $X(t_0)$: each of the initial conditions gives rise to a different bounded

solution of the LDE. If the uniqueness of the solution is to be preserved, the correct statement would be as follows:

Suppose that $(A(\cdot), B(\cdot))$ is uniformly stabilizable and $A(\cdot)$ and $B(\cdot)$ are bounded. Then, if there exists a reversed finite-time SPS solution $X(\cdot)$ of the LDE (2) on $(-\infty, t_0]$, $A(\cdot)$ is exponentially stable on $(-\infty, t_0]$. Conversely, if $A(\cdot)$ is exponentially stable on $(-\infty, t_0]$, there exists a unique reversed finite-time SPS solution $X(\cdot)$ of the LDE on $(-\infty, t_0]$.

Then, by taking the limit for $t_0 \rightarrow \infty$, Lemma 1 is obtained.

Finally, note that, if $A(\cdot)$ is exponentially stable, the infinite-time solution of the LDE proves to be a moving equilibrium for the set of all SPS finite-time solutions.

3. Maximal and nonnegative solutions. The present section will be devoted to the analysis of the nonnegative definite solutions of the RDE under the hypotheses of stabilizability and detectability, as well as detectability alone. We begin with two basic definitions.

DEFINITION 8 (maximal solution). A symmetric infinite-time solution $X^+(\cdot)$ of the RDE is said to be maximal if, for any symmetric infinite-time solution $X(\cdot)$ of the RDE, $X^+(t) - X(t) \geq 0$, for all t .

DEFINITION 9 (stabilizing solution). A *stabilizing* solution is an infinite-time symmetric solution $X(\cdot)$ of the RDE such that the corresponding closed-loop matrix $F(\cdot)$ is exponentially stable on $(-\infty, +\infty)$. A *finite-time stabilizing* solution is a symmetric finite-time solution $X(\cdot)$ of the RDE on $[t_0, +\infty)$ such that the corresponding closed-loop matrix $F(\cdot)$ is exponentially stable on $[t_0, +\infty)$.

Below, a sufficient condition for the existence of the maximal solution will be proved. Such a condition can be seen as an extension of a time-invariant result [7], [8]. However, differently from [7] and [8], herein the demonstration calls for a quasi-linearization (Newton) technique. The application of the Newton algorithm to the ARE traces back to [26], [27] for the continuous-time case, and [5], [28] for the discrete-time one. In all these papers, controllability and observability assumptions (sometimes relaxed to stabilizability and detectability) were made, so as to ensure the existence of a unique SPS solution of the ARE. In [6] it was pointed out that, under the sole hypothesis of observability, the Newton algorithm converges to the maximal (minimal) solution of the continuous-time ARE whenever it is started with a stabilizing (antistabilizing) gain. As for the time-varying case, we refer to [29], where, under the hypotheses of uniform reachability and uniform observability that guarantee the existence of a unique SPS moving equilibrium, the convergence of the quasi-linearization technique to such an equilibrium is proved. Finally, in a recent paper on the periodically time-varying difference Riccati equation [22], the sole-detectability has been shown to be a sufficient condition for the convergence of the Newton method to the periodic maximal solution. We generalize this result to the time-varying case by means of the following result.

THEOREM 1. *Let $(A(\cdot), C(\cdot))$ be uniformly detectable and consider the sequence of LDEs*

$$(4) \quad X_{i+1}(t+1) = A_i(t)X_{i+1}(t)A_i(t)' + Q(t) + K_i(t)R(t)K_i(t)', \quad i \geq 0,$$

where

$$A_i(t) = A(t) - K_i(t)C(t), \quad i \geq 0,$$

$$K_i(t) = A(t)X_i(t)C(t)'[R(t) + C(t)X_i(t)C(t)']^{-1}, \quad i \geq 1,$$

and $K_0(\cdot)$ is chosen so as to ensure the exponential stability of $A_0(\cdot)$. Then,

- (i) For each $i \geq 0$, (4) admits a unique real symmetric solution $X_{i+1}(\cdot)$. Moreover, this solution is SPS and such that $A_{i+1}(\cdot)$ is exponentially stable;
- (ii) For any t , $\{X_i(t)\}$ is a nonincreasing sequence that, for $i \rightarrow \infty$, converges to $\bar{X}(\cdot)$, where $\bar{X}(\cdot)$ is an infinite-time solution of the RDE (3);
- (iii) $\bar{X}(\cdot)$ coincides with the maximal solution $X^+(\cdot)$.

Proof. The proof of points (i) and (ii) is completely analogous to the proof of [22, Thm. 3], where a parallel result is established for periodic systems. The only difference is the use of the time-varying extended Lyapunov lemma in place of its periodic version. As for point (iii), consider any symmetric solution $X(\cdot)$ of the RDE and let $Y_i(t) = X_{i+1}(t) - X(t)$. Then, from (3) and (4) it can be seen that $Y_i(\cdot)$ is a symmetric solution of the following LDE:

$$Y_i(t+1) = A_i(t)Y_i(t)A_i(t)' + [K_i(t) - K(t)][C(t)X(t)C(t)' + R(t)][K_i(t) - K(t)]'.$$

In view of point (i), $A_i(\cdot)$ is exponentially stable, so that the extended Lyapunov lemma implies that $Y_i(t)$ is positive semidefinite for each t . Consequently, $\bar{X}(t) - X(t) = \lim_{i \rightarrow \infty} Y_i(t)$ is also positive semidefinite for each t . \square

The attention is now focused on the stabilizing solution and its relationships with the maximal one. If the RDE admits a stabilizing solution, we can use the corresponding Kalman gain to initialize the quasi-linearization procedure. This straightforwardly leads to the following result that, with reference to the infinite-dimensional and continuous-time case, can also be found in [18].

PROPOSITION 1. *The stabilizing solution of the RDE (3) (if any) coincides with the maximal solution and (therefore) is unique.*

Detectability is clearly a necessary condition for the existence of a stabilizing solution, but it is not sufficient. For example, in the theory of the time-invariant RDE, we must add the hypothesis that there are no (A, B) -unreachable eigenvalues on the unit circle to have a sufficient condition; see, e.g., [11], [12]. Although analogous sufficient conditions are available for the periodically time-varying case [21], the extension of this kind of result to the most general time-varying RDE appears difficult. However, it is still possible to provide a more restrictive sufficient condition that relies on detectability and stabilizability. The analysis of the Riccati equation under these hypotheses dates back to the papers by Wonham [3] and Kucera [4] on the ARE. The main result states that stabilizability and detectability constitute a necessary and sufficient condition for the existence of a unique SPS solution of the ARE and the stability of the closed-loop matrix. Moreover, such a solution of the ARE was shown to be attractive for all the SPS solutions of the associated RDE. When passing to the time-varying case, we must recall a result of Kalman [1]: uniform reachability and uniform observability guarantee the existence of a unique moving equilibrium, which is, in fact, stabilizing. The relaxation of the assumptions to uniform stabilizability and detectability is not difficult, thanks to the following result, which was proved in [17].

THEOREM 2 (Anderson and Moore [17]). *If $(A(\cdot), C(\cdot))$ is uniformly detectable and $(A(\cdot), Q(\cdot))$ is uniformly stabilizable, then every SPS (finite-time as well infinite-time) solution $X(\cdot)$ of the RDE is stabilizing.*

We are now in a position to give a sufficient condition for the existence of the stabilizing solution, as well for its attractiveness.

THEOREM 3. *If $(A(\cdot), C(\cdot))$ is uniformly detectable and $(A(\cdot), Q(\cdot))$ is uniformly stabilizable, then the RDE admits a unique SPS solution, which is, in fact, stabilizing. Moreover, such a solution is a moving equilibrium for all the SPS finite-time solutions of the RDE.*

Proof. The stabilizing property follows directly from Theorem 2. Any SPS solution being stabilizing, Proposition 1 implies that the SPS solution is unique and maximal. We now prove that the solution $X^+(\cdot)$ is a moving equilibrium. Indeed, let $X(\cdot)$ be any SPS finite-time solution of the RDE. It can be seen that

$$X^+(t+1) - X(t+1) = F^+(t)[X^+(t) - X(t)]F(t)',$$

where $F^+(\cdot)$ and $F(\cdot)$ are the closed-loop matrices relative to $X^+(\cdot)$ and $X(\cdot)$, respectively. Since $F^+(\cdot)$ and $F(\cdot)$ are exponentially stable, $X^+(t) - X(t)$ asymptotically goes to zero for $t \rightarrow \infty$. Indeed, denoting by $\Phi_{F^+}(t, t_0)$ and $\Phi_F(t, t_0)$ the transition matrix of $F^+(t)$ and $F(t)$, respectively, it follows that

$$\|X^+(t) - X(t)\| \leq \|\Phi_{F^+}(t, t_0)\| \|X^+(t_0) - X(t_0)\| \|\Phi_F(t, t_0)'\|,$$

where both $\|\Phi_{F^+}(t, t_0)\|$ and $\|\Phi_F(t, t_0)\|$ tend to zero for $t \rightarrow \infty$. \square

Finally, we will characterize the moving equilibrium property of the maximal solution in the nonstabilizable case. As seen in the previous theorem, the convergence analysis of the solutions of the RDE under stabilizability and detectability is made easy by the fact that every (finite-time or infinite-time) solution is stabilizing. Conversely, the difficulties in the nonstabilizable case are witnessed by the fact that, until recently [11], [12], there had been no systematic study even for the time-invariant RDE. Theorem 4, below, extends to the time-varying RDE a result relative to the stationary case, which can be found in [12, Thm. 4.2]. In the proof, the following lemma will be needed.

LEMMA 2. *Consider two RDEs (3) with the same $A(\cdot)$, $C(\cdot)$, and $R(\cdot)$ matrices but possibly different $Q(\cdot)$ matrices and possibly different initial conditions. Let the solutions to these RDEs be written as*

$$\begin{aligned} X_i(t+1) &= A(t)X_i(t)A(t)' + Q_i(t) \\ &\quad - A(t)X_i(t)C(t)'[C(t)X_i(t)C(t)' + R(t)]^{-1}C(t)X_i(t)A(t)', \end{aligned}$$

with $X_i(t_0) = X_{i,0}$, $i = 1, 2$. Then, $X_{1,0} \geq X_{2,0}$ and $Q_1(t) \geq Q_2(t)$, $t \geq t_0$, imply that $X_1(t) \geq X_2(t)$, $t \geq t_0$.

Proof. The proof is completely analogous to the proof given in [30] for the time-invariant version of the same lemma, and is therefore omitted. Note that the proof in [30] relies on a result originally proved by Nishimura [31] in a time-varying context.

THEOREM 4. *Assume that $(A(\cdot), C(\cdot))$ is uniformly detectable and let $X^+(\cdot)$ denote the maximal solution. Then, if $X(\cdot)$ is a finite-time solution of the RDE with initial condition $X(t_0) = X_0 \geq X^+(t_0)$, $\lim_{t \rightarrow \infty} X(t) - X^+(t) = 0$.*

Proof. The proof is inspired by [12], where an analogous result is proved in the time-invariant case. Consider the family of RDEs

$$\begin{aligned} X_k(t+1) &= A(t)X_k(t)A(t)' + Q_k(t) \\ &\quad - A(t)X_k(t)C(t)'[C(t)X_k(t)C(t)' + R(t)]^{-1}C(t)X_k(t)A(t)', \\ (5) \quad X_k(t_0) &= X_0, \\ Q_k(t) &= Q(t) + \frac{1}{k}I, \quad k = 1, 2, \dots \end{aligned}$$

Then, in view of Lemma 2, we have

$$X^+(t) \leq X(t) \leq X_{k+1}(t) \leq X_k(t), \quad t \geq t_0, \quad k = 1, 2, \dots$$

By the definition of $Q_k(t)$, $(A(\cdot), Q_k(\cdot))$ is uniformly stabilizable. Then, by Theorem 3, $\lim_{t \rightarrow \infty} X_k(t) = X_k^+(t)$, where $X_k^+(\cdot)$ is the maximal solution of (5). It is not difficult to see that

$$X^+(t) \leq X_{k+1}^+(t) \leq X_k^+(t) \quad \forall t, \quad k = 1, 2, \dots$$

Now, we take the limit for $k \rightarrow \infty$. $X_k^+(t)$ is monotonically nonincreasing (in the sense that $X_{k+1}^+(t) \leq X_k^+(t)$) and bounded below by $X^+(t)$. Hence, there exists some $\bar{X}(\cdot)$ such that $\lim_{k \rightarrow \infty} X_k^+(t) - \bar{X}(t) = 0$, for all t . Recalling (5), such a limit $\bar{X}(\cdot)$ proves to be an infinite-time solution of the RDE (3). Moreover, $\bar{X}(t) \geq X^+(t)$, for all t . Therefore, $X^+(\cdot)$ being maximal, $\bar{X}(\cdot) = X^+(\cdot)$. It follows easily that $\lim_{t \rightarrow \infty} X(t) - X^+(t) = 0$. \square

An analogous convergence result can be found in [18, Prop. 3.2], where the attractiveness from above of the stabilizing solution is established. Apart from the fact that [18] refers to a different context (infinite-dimensional, continuous-time, Riccati equations), Theorem 4 is therefore more general, in that the stabilizing solution, if any, is maximal, but the maximal solution is not necessarily stabilizing.

4. Minimal and nonpositive solutions. In this section, the nonpositive solutions of the RDE will be considered. The aim is to derive a set of results parallel to those established in the previous section. The key notions will be antidetectability and antistabilizability, as well as the notions of minimal and antistabilizing solution that are given below.

DEFINITION 10 (minimal solution). A symmetric solution $X^-(\cdot)$ of the RDE (3) is said to be minimal if, for any symmetric solution $X(\cdot)$ of the RDE, $X^-(t) - X(t) \leq 0$, for all t .

DEFINITION 11 (antistabilizing solution). A symmetric solution $X(\cdot)$ of the RDE is said to be antistabilizing if the corresponding closed-loop matrix function $F(\cdot)$ is exponentially antistable.

The analysis will be carried out by showing that there is a one-to-one correspondence between the nonpositive solutions of the RDE (3) and the nonnegative solutions of a suitably redefined RDE. In the continuous-time case, such a correspondence would be easily established by simply reversing the time axis. Given a Riccati differential equation characterized by $(A(t), Q(t), C(t), R(t))$, consider the “reversed” Riccati differential equation corresponding to $(-A(-t), Q(-t), C(-t), R(-t))$: if $X(t)$ is an SNS solution of the original Riccati equation, $-X(-t)$ is an SPS solution of the “reversed” equation, and vice versa. In discrete time, as shown below, the correspondence is not so easily worked out, and some technicalities are required. From now on, it is implicitly assumed that $A(\cdot)^{-1}$ exists and is bounded. Some discussion on the merits of such assumption can be found at the end of this section. First, let us define the “reversed system.”

DEFINITION 12 (reversed system). Assume that $A(t)$ is nonsingular for each t and let

$$(6a) \quad \tilde{A}(t) = A(-t)^{-1},$$

$$(6b) \quad \tilde{Q}(t) = A(-t)^{-1}Q(-t)A(-t)^{-1},$$

$$(6c) \quad \tilde{C}(t) = C(-t+1),$$

$$(6d) \quad \tilde{R}(t) = R(-t+1).$$

Then the reversed system associated with system (1) is

$$(7a) \quad \tilde{x}(t+1) = \tilde{A}(t)\tilde{x}(t) + \tilde{v}(t),$$

$$(7b) \quad \tilde{y}(t) = \tilde{C}(t)\tilde{x}(t) + \tilde{w}(t),$$

where $\tilde{v}(\cdot)$ and $\tilde{w}(\cdot)$ are independent zero-mean white noises having covariance matrices $\tilde{Q}(\cdot)$ and $\tilde{R}(\cdot)$, respectively.

Observe that, if $\tilde{x}(\tau) = x(-\tau + 1)$, $\tilde{v}(t) = -\tilde{A}(t)v(-t)$, and $\tilde{w}(t) = w(-t + 1)$, $t \geq \tau$, then $\tilde{x}(t) = x(-t + 1)$ and $\tilde{y}(t) = y(-t + 1)$, for each $t \geq \tau$. In other words, the reversed system is nothing but a representation of system (1) when the time axis is reversed. Note also that by reversing system (7), we turn back to the original system (1).

Associated with the reversed system (7) is the following reversed Riccati difference equation:

$$(8) \quad \begin{aligned} \tilde{X}(t+1) &= \tilde{A}(t)\tilde{X}(t)\tilde{A}(t)' + \tilde{Q}(t) \\ &\quad - \tilde{A}(t)\tilde{X}(t)\tilde{C}(t)'[\tilde{C}(t)\tilde{X}(t)\tilde{C}(t)' + \tilde{R}(t)]^{-1}\tilde{C}(t)\tilde{X}(t)\tilde{A}(t)'. \end{aligned}$$

To clarify the relationships between the solutions of (8) and (3), we first provide a result on the structural properties of (7). The proof is straightforward and is therefore left to the reader.

PROPOSITION 2.

- (a) Matrix $\tilde{A}(\cdot)$ is exponentially stable (antistable) if and only if $A(\cdot)$ is exponentially antistable (stable).
- (b) The pair $(\tilde{A}(\cdot), \tilde{Q}(\cdot))$ $[(\tilde{A}(\cdot), \tilde{C}(\cdot))]$ is uniformly reachable (observable) if and only if the pair $(A(\cdot), Q(\cdot))$ $[(A(\cdot), C(\cdot))]$ is uniformly reachable (observable).
- (c) The pair $(\tilde{A}(\cdot), \tilde{Q}(\cdot))$ is uniformly stabilizable (antistabilizable) if and only if the pair $(A(\cdot), Q(\cdot))$ is uniformly antistabilizable (stabilizable).
- (d) The pair $(\tilde{A}(\cdot), \tilde{C}(\cdot))$ is uniformly detectable (antidetectable) if and only if the pair $(A(\cdot), C(\cdot))$ is uniformly antidetectable (detectable).

We are now in a position to prove the following theorem that establishes a one-to-one correspondence between the SPS (SNS) solutions of (8) and the SNS (SPS) solutions of (3).

THEOREM 5. Assume that $A(t)$ is nonsingular for each t . Let $Y(-t + 1) = -X(t)$, and

$$(9) \quad \tilde{X}(t+1) = \tilde{A}(t)Y(t)\tilde{A}(t)' + \tilde{Q}(t).$$

Then

- (a) $\tilde{X}(\cdot)$ is an SPS (SNS) solution of the RDE (8) if and only if $X(\cdot)$ is an SNS (SPS) solution of the RDE (3);
- (b) Denoting by $K(\cdot)$ the Kalman gain associated with an SNS (SPS) solution $X(\cdot)$ of the RDE (3) and letting $\tilde{K}(t) = \tilde{A}(t)\tilde{X}(t)\tilde{C}(t)'[\tilde{C}(t)\tilde{X}(t)\tilde{C}(t)' + \tilde{R}(t)]^{-1}$, the closed-loop matrix $\tilde{A}(\cdot) - \tilde{K}(\cdot)\tilde{C}(\cdot)$ is antistable (stable) if and only if the closed-loop matrix $\tilde{A}(\cdot) - \tilde{K}(\cdot)\tilde{C}(\cdot)$ is stable (antistable);
- (c) Letting $X_1(\cdot)$, $X_2(\cdot)$ be two SNS (SPS) solutions of the RDE (3) and $\tilde{X}_1(\cdot)$, $\tilde{X}_2(\cdot)$, the corresponding SPS (SNS) solutions of the RDE (8), $X_1(\cdot) \geq X_2(\cdot)$ if and only if $\tilde{X}_1(\cdot) \leq \tilde{X}_2(\cdot)$.

Proof of (a). Assume that $X(\cdot)$ is an SNS solution of the RDE (3). Then

$$\begin{aligned} -Y(t) &= -\tilde{A}(t)^{-1}Y(t+1)\tilde{A}(t)^{-1} + \tilde{A}(t)^{-1}\tilde{Q}(t)\tilde{A}(t)^{-1} \\ &\quad - \tilde{A}(t)^{-1}Y(t+1)\tilde{C}(t+1)'[\tilde{R}(t+1) - \tilde{C}(t+1)Y(t+1)\tilde{C}(t+1)']^{-1} \\ &\quad \cdot \tilde{C}(t+1)Y(t+1)\tilde{A}(t)^{-1}, \end{aligned}$$

which implies that

$$(10) \quad \tilde{X}(t) = Y(t) + Y(t)\tilde{C}(t)'[\tilde{R}(t) - \tilde{C}(t)Y(t)\tilde{C}(t)']^{-1}\tilde{C}(t)Y(t).$$

Now, in view of Lemma A1 (in the Appendix), $\tilde{R}(t) - \tilde{C}(t)Y(t)\tilde{C}(t)' > 0$. Therefore, from (10) and Lemma A2, it follows that

$$(11) \quad Y(t) = \tilde{X}(t) - \tilde{X}(t)\tilde{C}(t)'[\tilde{R}(t) + \tilde{C}(t)\tilde{X}(t)\tilde{C}(t)']^{-1}\tilde{C}(t)\tilde{X}(t).$$

By substituting this expression in (9), it is easy to see that $\tilde{X}(\cdot)$ satisfies the RDE (8).

Conversely, suppose that $\tilde{X}(\cdot)$ is an SPS solution of the RDE (8). From (8) and (9), it follows that (11) holds. Now, in view of Lemma A3, (10) holds, also, so that by substituting (10) into (9), $X(\cdot)$ is easily shown to satisfy the RDE (3).

As for the SPS solutions of the RDE (3), recall that the reversed system associated with (7) coincides with system (1). Therefore, there also exists a one-to-one correspondence between the SPS solutions of (3) and the SNS solutions of (8).

Proof of (b). Let us consider an SNS solution $X(\cdot)$ of (3). First, note that the invertibility of $A(t)$ and Lemma A4 imply that $\tilde{A}(t) - \tilde{K}(t)\tilde{C}(t)$ is nonsingular, for all t . By resorting to the matrix inversion lemma,

$$\tilde{A}(t) - \tilde{K}(t)\tilde{C}(t) = [[I + \tilde{X}(t)\tilde{C}(t)'\tilde{R}(t)^{-1}\tilde{C}(t)]\tilde{A}(t)^{-1}]^{-1}.$$

Define $\Gamma(-t+1) = [I + \tilde{X}(t)\tilde{C}(t)'\tilde{R}(t)^{-1}\tilde{C}(t)]$. Note that, since $\tilde{X}(\cdot)$, $\tilde{C}(\cdot)$, and $\tilde{R}(\cdot)^{-1}$ are bounded, $\Gamma(\cdot)^{-1}$ is bounded. As seen above, expression (10) holds true. Then, the matrix inversion lemma implies that

$$\begin{aligned} \tilde{X}(t)\tilde{C}(t)'\tilde{R}(t)^{-1}\tilde{C}(t) &= [Y(t) + Y(t)\tilde{C}(t)'\tilde{R}(t) - \tilde{C}(t)Y(t)\tilde{C}(t)']^{-1}\tilde{C}(t)Y(t)\tilde{C}(t)'\tilde{R}(t)^{-1}\tilde{C}(t) \\ &= Y(t)\tilde{C}(t)'\tilde{R}(t) - \tilde{C}(t)Y(t)\tilde{C}(t)']^{-1}\tilde{C}(t). \end{aligned}$$

Therefore,

$$\begin{aligned} \Gamma(t) &= I - X(t)C(t)'\tilde{R}(t) + C(t)X(t)C(t)']^{-1}C(t) \quad \text{and} \\ \tilde{A}(t) - \tilde{K}(t)\tilde{C}(t) &= A(-t)^{-1}\Gamma(-t+1)^{-1}. \end{aligned}$$

Denote by $\Psi(k, s)$ and $\tilde{\Psi}(k, s)$ the transition matrices relative to $A(\cdot) - K(\cdot)C(\cdot)$ and $\tilde{A}(\cdot) - \tilde{K}(\cdot)\tilde{C}(\cdot)$, respectively. Then, $\tilde{\Psi}(k, s) = A(-k+1)^{-1}\Psi(-s, -k+2)^{-1}\Gamma(-s+1)^{-1}$, and the thesis is proved.

As for the SPS solutions of the RDE (3), the same observation as at the end of the proof of (a) applies.

Proof of (c). The proof is straightforward in view of (9). \square

When $X(\cdot)$ is SNS, $\tilde{X}(\cdot)$ defined in (9) is SPS and satisfies the RDE (8). Hence, $\tilde{X}(\cdot)$ can be seen as the covariance of the state prediction error relative to system (7). Now, (9) is the classical time-update equation of the Kalman filter, so that $Y(\cdot)$ coincides with the covariance of the state filtering error. Therefore, given any SNS solution $X(\cdot)$ of (3), its reversed opposite $Y(\cdot)$ can be interpreted as the covariance of a state *filtering* error relative to the reversed system (7). On the other hand, the SPS solutions are, as usual, interpreted as the covariance of a state *prediction* error relative to system (1). This is a main difference with respect to the continuous-time case, where both the SPS solutions and the reversed opposite of the SNS solutions of the Riccati differential equation can be seen as the covariance of a state *filtering* error relative to a suitable (standard or reversed) system.

Now, in view of Theorem 5 and Proposition 2, the following results on the SNS solutions of (3) are direct consequences of the parallel results on the SPS solutions that were proved in the previous section.

THEOREM 6. *Let $(A(\cdot), C(\cdot))$ be uniformly antidetactable. Then the RDE (3) admits a minimal solution $X^-(\cdot)$, which is, in fact, negative semidefinite.*

PROPOSITION 3. *The antistabilizing solution of the RDE (3) (if any) coincides with the minimal solution and is unique.*

THEOREM 7. *If $(A(\cdot), C(\cdot))$ is uniformly antidetactable and $(A(\cdot), Q(\cdot))$ is uniformly antistabilizable, then the RDE (3) admits a unique SNS solution, which is, in fact, antistabilizing. Moreover, such a solution is a moving equilibrium for all the SNS reversed finite-time solutions of the RDE.*

THEOREM 8. Assume that $(A(\cdot), C(\cdot))$ is uniformly antidetectable and let $X^-(\cdot)$ denote the minimal solution. Then, if $X(\cdot)$ is a reversed finite-time solution of the RDE with initial condition $X(t_0) = X_0 \leq X^-(t_0)$, $\lim_{t \rightarrow -\infty} X(t) - X^-(t) = 0$.

Remark 3. We wonder whether the assumption on the reversibility of $A(\cdot)$ is really necessary to establish the above results on the existence and attractiveness of the minimal and antistabilizing solution. In fact, as shown below by means of a counterexample, if the reversibility assumption is removed, the minimal solution may not even exist.

Counterexample. Let $Q(t) = 0$, $C(t) = 1$, $R(t) = 1$, for all t , and

$$A(t) = \begin{cases} 1, & t < 0, \\ 0, & t \geq 0. \end{cases}$$

Correspondingly, the general expression for the SNS infinite-time solutions of the RDE is

$$X(t) = \begin{cases} \frac{X_0}{1+tX_0}, & t \leq 0, \\ 0, & t \geq 1, \end{cases}$$

where X_0 is an arbitrary negative scalar. Hence, it is apparent that, in spite of the uniform observability of $(A(\cdot), C(\cdot))$, there does not exist a minimal solution. Moreover, no infinite-time solution is antistabilizing.

5. Lattice of the solutions. In this section we will characterize the set of all real symmetric infinite-time solutions of the time-varying RDE. If we restrict our attention to the constant solutions of the time-invariant RDE, a consistent and comprehensive theory, at least for the continuous time case, is already available; see, e.g., [6], [8], [13], [10], [14]. The most remarkable result is perhaps the one that states that if (A, C) is observable, the set of all real symmetric solutions of the ARE is a complete lattice with respect to the ordering of symmetric matrices. The primary aim of this section will consist in extending such a result to the time-varying case. To this purpose, we will resort to a device (Lemma 3), which allows us to reduce the analysis of all the symmetric solutions of the RDE to the analysis of the SPS solutions of a suitably redefined RDE. As by-products of our analysis, some results concerning the convergence properties and the gap between the maximal and minimal solution will also be obtained. Throughout this section, it will be assumed that $A(\cdot)^{-1}$ exists and is bounded.

LEMMA 3. Let $\bar{X}(\cdot)$ be a (finite-time or infinite-time) solution of the RDE (3) and consider the new RDE

$$(12) \quad \begin{aligned} Z(t+1) = & \bar{F}(t)Z(t)\bar{F}(t)' - \bar{F}(t)Z(t)C(t)'[C(t)Z(t)C(t)' \\ & + \bar{R}(t)]^{-1}C(t)Z(t)\bar{F}(t)', \end{aligned}$$

where

$$\begin{aligned} \bar{R}(t) &= C(t)\bar{X}(\cdot)C(t)' + R(t), \\ \bar{F}(t) &= A(t) - \bar{K}(t)C(t), \\ \bar{K}(t) &= A(t)\bar{X}(t)C(t)'[C(t)\bar{X}(t)C(t)' + R(t)]^{-1}. \end{aligned}$$

Then, $Z(\cdot)$ is a solution of (12) if and only if $X(\cdot) = Z(\cdot) + \bar{X}(\cdot)$ is a solution of (3). Moreover, the closed-loop matrix relative to $Z(\cdot)$ coincides with the closed-loop matrix relative to $X(\cdot)$, shown below:

$$\begin{aligned} & \bar{F}(t) - \bar{F}(t)Z(t)C(t)'[C(t)Z(t)C(t)' + \bar{R}(t)]^{-1}C(t) \\ &= A(t) - A(t)X(t)C(t)'[C(t)X(t)C(t)' + R(t)]^{-1}C(t) \quad \forall t. \end{aligned}$$

Finally, $(\bar{F}(\cdot), C(\cdot))$ is uniformly detectable (uniformly observable) if and only if $(A(\cdot), C(\cdot))$ is uniformly detectable (uniformly observable).

Proof. The first point straightforwardly follows from [32, Lemma 3.1], and the second one is verified by inspection. As for detectability and observability, recall that they are feedback invariant properties; see, e.g., [25] and [17]. \square

We wonder whether, when passing from (3) to (12), the properties concerning the boundedness of $A(\cdot)$ translate into analogous properties relative to $F(\cdot)$. The answer can be found in the following result.

LEMMA 4. Assume that $A(\cdot)$ and $A(\cdot)^{-1}$ are bounded. Consider the closed-loop matrix $F(\cdot)$ corresponding to any infinite-time solution $X(\cdot)$. Then, $F(\cdot)$ and $F(\cdot)^{-1}$ are bounded, too.

Proof. By means of the matrix inversion lemma, we can see that

$$F(t) = A(t) - K(t)C(t) = A(t)[I + X(t)C(t)'R(t)^{-1}C(t)].$$

Then, the thesis easily follows from the boundedness of $X(\cdot)$, $C(\cdot)$, and $R(\cdot)^{-1}$. \square

By means of Lemma 3, in the case where a minimal solution $X^-(\cdot)$ exists, we can take $\bar{X}(\cdot) = X^-(\cdot)$, so that any infinite-time solution of the RDE (3) corresponds to a positive semidefinite solution of (12). Then, when $X^-(\cdot)$ is antistabilizing, the RDE (12) is just a particular case of the RDE (3) with $A(\cdot)$ exponentially antistable and $Q(t) = 0$. A result concerning the maximal solution of such a particular RDE is now established.

LEMMA 5. Assume that $A(\cdot)$ is exponentially antistable, $Q(t) = 0$, for all t , and $(A(\cdot), C(\cdot))$ is uniformly detectable. Then the maximal solution $X^+(\cdot)$ of the RDE (3) is positive definite, and $X^+(\cdot)^{-1}$ is bounded.

Proof. By Theorem 4, a finite-time solution $X(\cdot)$ of the RDE with initial condition $X(t_0) \geq X^+(t_0)$ converges to the maximal solution. In particular, assume that $X(t_0) > 0$. We will prove that the maximal solution is positive definite by showing that $X(t)^{-1}$ remains bounded for each t . Consider the RDE (3) with $Q(t) = 0$, for all t . By means of the matrix inversion lemma, we obtain

$$\begin{aligned} X(t+1) &= A(t)[X(t)^{-1} + C(t)'R(t)^{-1}C(t)]^{-1}A(t)', \\ (13) \quad X(t+1)^{-1} &= A(t)'^{-1}X(t)^{-1}A(t)^{-1} + A(t)'^{-1}C(t)'R(t)^{-1}C(t)A(t)^{-1}. \end{aligned}$$

Recall that the exponential antistability of $A(\cdot)$ entails the exponential stability of $A(\cdot)^{-1}$. Since $A(\cdot)^{-1}$, $C(\cdot)$, and $R(\cdot)^{-1}$ are bounded, $X(\cdot)^{-1}$ is bounded, also. \square

Interestingly enough, this last result enables us to clarify some relationships between the maximal and the minimal solution. In particular, we focus on the so-called gap between these solutions and on their stabilizing and convergence properties.

COROLLARY 1. Assume that $(A(\cdot), C(\cdot))$ is uniformly observable and that the minimal solution is antistabilizing. Then the gap $X^+(t) - X^-(t)$ between the maximal and the minimal solution of the RDE is positive definite for each t .

Proof. With reference to Lemma 3, let $\bar{X}(\cdot) = X^-(\cdot)$. By the assumptions, $\bar{F}(\cdot)$ is exponentially antistable. Then, the gap $X^+(\cdot) - X^-(\cdot)$ turns out to be the maximal solution of the RDE (12), and the result follows from Lemma 5.

LEMMA 6. The maximal solution of the RDE (3) is stabilizing if and only if the minimal solution is antistabilizing.

Proof. Suppose that the minimal solution is antistabilizing. In view of Lemma 3, there is no loss of generality in assuming that $Q(t) = 0$, for all t , and $A(\cdot)$ is exponentially antistable. Then, by definition of infinite-time solution and by Lemma 5, the maximal solution $X^+(\cdot)$, as well as its inverse, is bounded. Under the given assumptions, the

RDE can be rewritten as

$$\begin{aligned} X^+(t) &= A(t)^{-1} X^+(t+1) A(t)^{-1'} \\ &\quad + A(t)^{-1} K^+(t) [C(t) X^+(t) C(t)' + R(t)] K^+(t)' A(t)^{-1'}. \end{aligned}$$

By applying Lemma A5, we can see that the pair $(A(\cdot)^{-1}, A(\cdot)^{-1} K^+(\cdot))$ is uniformly reachable. Then, Proposition 2(b) implies the uniform reachability of $(A(\cdot), K^+(\cdot))$. By recalling that reachability is feedback invariant [25], this, in turn, leads to the uniform reachability of $(A(\cdot) - K^+(\cdot) C(\cdot), K^+(\cdot))$. The RDE can be given the following expression:

$$X^+(t+1) = F^+(t) X^+(t) F^+(t)' + K^+(t) R(t) K^+(t)',$$

where $F^+(\cdot) = A(\cdot) - K^+(\cdot) C(\cdot)$ is the closed-loop matrix relative to $X^+(\cdot)$. Since the pair $(F^+(\cdot), K^+(\cdot))$ is uniformly reachable, the exponential stability of $F^+(\cdot)$ follows from the Lyapunov lemma applied to this last equation. The only "if" part of the proof is a straightforward consequence of the properties of the reversed Riccati equation.

LEMMA 7. Assume that $(A(\cdot), C(\cdot))$ is uniformly observable and the minimal solution is antistabilizing. Consider a finite-time (reversed finite-time) solution $X(\cdot)$ of the RDE (3) with initial condition $X(t_0) > X^-(t_0)$ [$X(t_0) < X^+(t_0)$]. Then $\lim_{t \rightarrow \infty} X(t) - X^+(t) = 0$ [$\lim_{t \rightarrow -\infty} X(t) - X^-(t) = 0$].

Proof. Let $\bar{X}(\cdot) = X^-(\cdot)$. Then, by Lemma 3, if $X(\cdot)$ is a finite-time solution of (3), $Z(\cdot) = X(\cdot) - X^-(\cdot)$ satisfies (12) and $Z(t_0) > 0$. Therefore, we can assume, without any loss of generality, that $A(\cdot)$ is exponentially antistable, $Q(t) = 0$, for all t , and $X(t_0) > 0$. Now, following the proof of Lemma 5, we can easily see that $X^+(\cdot)^{-1}$ is the unique SPS solution of the LDE (13). Note that $A(\cdot)^{-1}$ is exponentially stable. Then, (13) implies that, independently of the initial condition, for $t \rightarrow \infty$, $X(t)^{-1}$ converges to $X^+(t)^{-1}$.

As for the reversed finite-time solution, the proof is completely analogous.

COROLLARY 2. Assume that $(A(\cdot), C(\cdot))$ is uniformly observable and the minimal solution is antistabilizing. Let $X(\cdot)$ be a finite-time (reversed finite-time) solution such that $X(s) > 0$ [$X(s) < 0$], where s is an arbitrary time point. Then, $\lim_{t \rightarrow \infty} X(t) - X^+(t) = 0$ [$\lim_{t \rightarrow -\infty} X(t) - X^-(t) = 0$].

A definition is now given that plays a keyrole in the characterization of the symmetric solutions.

DEFINITION 13 (supporting subspace). Let $X(\cdot)$ be a (finite-time or infinite-time) solution of the RDE (3). Assume that there exists a minimal solution $X^-(\cdot)$ and let $Z(\cdot) = X(\cdot) - X^-(\cdot)$. Then, the time-varying subspace $\mathcal{R}[Z(\cdot)]$ is said to be the supporting subspace of $X(\cdot)$.

Obviously, the supporting subspace of the minimal solution is the origin. As for the maximal one, note that, under the assumptions of Corollary 1, its supporting subspace coincides with R^n .

In the following lemma a chain of results is developed to reach the main result of the section, which is stated in Theorem 9.

LEMMA 8. Let $X(\cdot)$ be an SPS (finite-time or infinite-time) solution of the RDE (3). Assume that $X(\cdot)$ is not identically equal to zero, $Q(t) = 0$, for all t , $A(\cdot)$ is exponentially antistable, and $(A(\cdot), C(\cdot))$ is uniformly observable. Then,

(a) If $X(t_0)$ is not positive definite, there exists a unitary matrix function $T(t) = T(t)^{-1'}$, defined on $[t_0, \infty)$ such that

$$(14) \quad T(t) X(t) T(t)' = \begin{bmatrix} X_1(t) & 0 \\ 0 & 0 \end{bmatrix}, \quad T(t+1) A(t) T(t)' = \begin{bmatrix} A_1(t) & * \\ 0 & A_2(t) \end{bmatrix}, \quad t \geq t_0,$$

where $A_1(t)$ and $X_1(t)$ are square matrices of the same constant dimensions, $X_1(t_0) > 0$, and $*$ denotes a term we do not consider specifically;

- (b) Partition matrix $C(t)T(t)'$ as $C(t)T(t)' = [C_1(t)*]$, where $C_1(t)$ has the same number of columns as $A_1(t)$. Then $X_1(t)$, $t \geq t_0$, satisfies the following reduced-order RDE:

$$(15) \quad \begin{aligned} X_1(t+1) &= A_1(t)X_1(t)A_1(t)' - A_1(t)X_1(t)C_1(t)' \\ &\quad \cdot [C_1(t)X_1(t)C_1(t)' + R(t)]^{-1}C_1(t)X_1(t)A_1(t)'; \end{aligned}$$

- (c) The dimensions of the supporting subspace of $X(\cdot)$ are time-invariant;
 (d) There exists a unitary $T(t)$ such that the decomposition (14) (with $X_1(t) > 0$) and expression (15) hold for any t ;
 (e) If $X(\cdot)$ is a finite-time solution, for $t \rightarrow \infty$, it converges to

$$(16) \quad X_\infty(\cdot) = T(\cdot)' \begin{bmatrix} X_1^+(\cdot) & 0 \\ 0 & 0 \end{bmatrix} T(\cdot),$$

where $X_1^+(\cdot)$ is the maximal solution of (15). Moreover, $X_\infty(\cdot)$ has the same supporting subspace as $X(\cdot)$;

- (f) Let $X_a(\cdot)$ and $X_b(\cdot)$ be two (finite-time or infinite-time) SPS solutions of the RDE and denote by $\mathcal{X}_a(\cdot)$ and $\mathcal{X}_b(\cdot)$ the corresponding supporting subspaces. Then, if $\mathcal{X}_a(s) = \mathcal{X}_b(s) [\mathcal{X}_a(s) \supseteq \mathcal{X}_b(s)]$ at an arbitrary time point s , $\mathcal{X}_a(t) = \mathcal{X}_b(t) [\mathcal{X}_a(t) \supseteq \mathcal{X}_b(t)]$, for any t ;
 (g) Let $X_a(\cdot)$ and $X_b(\cdot)$ be two infinite-time SPS solutions of the RDE. Then, $\mathcal{X}_a(t) = \mathcal{X}_b(t)$ implies $X_a(\cdot) = X_b(\cdot)$;
 (h) Given a time point s and a subspace \mathcal{X} of R^n , there exists one and only one infinite-time SPS solution $\tilde{X}(\cdot)$ such that $\mathcal{R}[\tilde{X}(s)] = \mathcal{X}$;
 (i) Let $X_a(\cdot)$ and $X_b(\cdot)$ be two infinite-time SPS solutions of the RDE, and denote by $\mathcal{X}_a(\cdot)$ and $\mathcal{X}_b(\cdot)$ the corresponding supporting subspaces. Then, if $\mathcal{X}_a(s) \supseteq \mathcal{X}_b(s)$ at an arbitrary time point s , $X_a(t) \supseteq X_b(t)$, for all t .

Proof of (a). First, note that $T(t)$ defines a change of basis for system (1). It can be seen that $X(\cdot)$ is a solution of the RDE (3) if and only if $\hat{X}(t) = T(t)X(t)T(t)'$ satisfies the RDE

$$(17) \quad \begin{aligned} \hat{X}(t+1) &= \hat{A}(t)\hat{X}(t)\hat{A}(t)' \\ &\quad - \hat{A}(t)\hat{X}(t)\hat{C}(t)'[\hat{C}(t)\hat{X}(t)\hat{C}(t)' + R(t)]^{-1}\hat{C}(t)\hat{X}(t)\hat{A}(t)', \end{aligned}$$

where $\hat{A}(t) = T(t+1)A(t)T(t)'$ and $\hat{C}(t) = C(t)T(t)'$.

Since $X(t_0)$ is SPS, there exists a unitary $T(t_0)$ such that

$$T(t_0)X(t_0)T(t_0)' = \begin{bmatrix} X_1(t_0) & 0 \\ 0 & 0 \end{bmatrix},$$

with $X_1(t_0)$ positive definite. Now, denoting by r the rank of $X(t_0)$, let the first r columns of $T(t_0+1)^{-1}$ be an orthonormal basis of $\mathcal{R}[A(t_0)X(t_0)]$ and choose the other columns so as to make $T(t_0+1)$ unitary. Then, $\hat{A}(t_0)$ has the structure given in (14). By substituting $\hat{A}(t_0)$ and $\hat{X}(t_0)$ into (17), $\hat{X}(t_0+1)$ also takes on the partitioned structure (14). Obviously, the procedure can be iterated for any $t \geq t_0$, proving the thesis.

Proof of (b). The proof follows by simple substitution of $\hat{A}(t)$, $\hat{C}(t)$, and $\hat{X}(t)$ into (17).

Proof of (c). Letting

$$F_1(t) = A_1(t) - K_1(t)C_1(t),$$

$$K_1(t) = A_1(t)X_1(t)C_1(t)'[C_1(t)X_1(t)C_1(t)' + R(t)]^{-1},$$

(15) entails that

$$(18) \quad X_1(t+1) = F_1(t)X_1(t)F_1(t)' + K_1(t)R(t)K_1(t)'.$$

Note that, in view of the particular structure of $\hat{A}(t)$, the nonsingularity of $A(t)$ implies the nonsingularity of $A_1(t)$. Since $X_1(t) > 0$, for all t , Lemma A4 implies that $F_1(t)$ is nonsingular, so that the thesis is a direct consequence of (18).

Proof of (d). It is a consequence of the invariance of the dimensions of the supporting subspace just proved in point (c).

Proof of (e). The uniform observability of the pair $(A(\cdot), C(\cdot))$ implies the observability of $(\hat{A}(\cdot), \hat{C}(\cdot))$, which, in turn, implies the uniform observability of $(A_1(\cdot), C_1(\cdot))$. Therefore, by Theorem 1, the RDE (15) admits a maximal solution $X_1^+(\cdot)$. Since $X_1(t_0) > 0$, Corollary 2 entails the convergence of $X_1(\cdot)$ to the maximal solution $X_1^+(\cdot)$. Finally, recall that $X(\cdot)$ is a solution of (3) if and only if $T(\cdot)\hat{X}(\cdot)T(\cdot)'$ is a solution of (17). As for the supporting subspaces, Lemma 5 implies that $X_1^+(t)$ is positive definite for each t , so that $\mathcal{R}[X_\infty(t)] = \mathcal{R}[X(t)]$, for all t .

Proof of (f). Assume that $\mathcal{X}_a(s) = \mathcal{X}_b(s)$. As seen in the proof of point (a), $\mathcal{R}[X(s+1)]$ depends only on $\mathcal{R}[X(s)]$ and not on the particular value of $X(s)$. Therefore, $\mathcal{X}_a(t) = \mathcal{X}_b(t)$, $t \geq s$. By making reference to the reversed RDE, we can see that also $\mathcal{R}[X(s-1)]$ depends only on $\mathcal{R}[X(s)]$, and the first part of the proof is completed.

As for $\mathcal{X}_a(s) \supseteq \mathcal{X}_b(s)$, consider the following nonsingular transformation $T_a(\cdot)$ that performs the decomposition (14) on $X_a(\cdot)$:

$$T_a(t)X_a(t)T_a(t)' = \begin{bmatrix} X_{a1}(t) & 0 \\ 0 & 0 \end{bmatrix}, \quad t \geq s,$$

with $X_{a1}(t) > 0$. It is easily seen that

$$T_a(t)X_b(t)T_a(t)' = \begin{bmatrix} X_{b1}(t) & 0 \\ 0 & 0 \end{bmatrix}, \quad t \geq s,$$

with $X_{b1}(t)$ of the same dimensions as $X_{a1}(t)$. Therefore, $\mathcal{X}_a(t) \supseteq \mathcal{X}_b(t)$, $t \geq s$. Analogously, by means of the reversed RDE, it can be shown that $\mathcal{X}_a(s) \supseteq \mathcal{X}_b(s)$ implies that $\mathcal{X}_a(t) \supseteq \mathcal{X}_b(t)$, $t < s$.

Proof of (g). In view of point (f), $\mathcal{X}_a(\cdot) = \mathcal{X}_b(\cdot)$. For both solutions, decomposition (14) can be performed by means of the same transformation $T(\cdot)$. Then, point (e) implies that both $X_a(\cdot)$ and $X_b(\cdot)$ coincide with $X_\infty(\cdot)$ defined in (16).

Proof of (h). Consider any finite-time solution $X(\cdot)$ such that $\mathcal{R}[X(s)] = \mathcal{X}$. Let $\bar{X}(\cdot) = X_\infty(\cdot)$ be the infinite-time solution to which, in view of point (e), $X(\cdot)$ converges. Then, by point (g), $\bar{X}(\cdot)$ is the only infinite-time solution such that $\mathcal{R}[\bar{X}(s)] = \mathcal{X}$.

Proof of (i). First, note that, by point (f), $\mathcal{X}_a(s) \supseteq \mathcal{X}_b(s)$ implies that $\mathcal{X}_a(t) \supseteq \mathcal{X}_b(t)$, for all t . By the assumptions, for any τ , it is possible to choose two initial conditions

$X_\alpha(\tau)$ and $X_\beta(\tau)$, such that $\mathcal{R}[X_\alpha(\tau)] = \mathcal{X}_a(\tau)$, $\mathcal{R}[X_\beta(\tau)] = \mathcal{X}_b(\tau)$ and $X_\alpha(\tau) \geq X_\beta(\tau)$. Denote by $X_\alpha(\cdot, \tau)$ and $X_\beta(\cdot, \tau)$ the finite-time solutions such that $X_\alpha(\tau, \tau) = X_\alpha(\tau)$, $X_\beta(\tau, \tau) = X_\beta(\tau)$. Lemma 2 implies that $X_\alpha(t, \tau) \geq X_\beta(t, \tau)$, $t \geq \tau$. Since $X_a(t) = \lim_{\tau \rightarrow -\infty} X_\alpha(t, \tau)$ and $X_b(t) = \lim_{\tau \rightarrow -\infty} X_\beta(t, \tau)$, the thesis follows.

THEOREM 9. *Let $(A(\cdot), C(\cdot))$ be uniformly observable and assume that the minimal solution of the RDE (3) is antistabilizing. Then, the infinite-time solutions of the RDE constitute an infinite number of isomorphic distributive lattices with common minimal and maximal elements.*

Proof. Consider the set of all subspaces in R^n . In view of Lemma 8(h), to each of these subspaces corresponds one and only one infinite-time solution. We now show that this infinite number of solutions can be organized in an infinity of isomorphic lattices. Take n independent one-dimensional subspaces $\mathcal{X}_i, i = 1, \dots, n$, and call $\mathcal{S} = \{\mathcal{X}_i\}$ the set of these subspaces. The set \mathcal{T} formed by all subspaces that can be obtained by means of the operations of intersection and sum between the elements of \mathcal{S} turns out to be a distributive lattice with respect to such operations. Note that, independently of the choice of \mathcal{S} , the origin and R^n belong to \mathcal{T} and constitute the maximal and minimal element of \mathcal{T} . By suitably varying \mathcal{S} , we can see that all the subspaces in R^n can be organized in an infinity of isomorphic lattices. Thanks to the one-to-one correspondence between subspaces and solutions of the RDE, this reflects in an analogous structure for the infinite-time solutions. The partial ordering by inclusion of the subspaces is translated into the partial ordering \geq for symmetric matrices (Lemma 8(i)), while the maximal and minimal elements of the lattices are given by the maximal and minimal solution.

Remark 4. Obviously, the above theorem also holds for the time-invariant RDE. Then, the problem of determining the number and the structure of the solutions of the ARE is equivalent to the problem of determining which infinite-time solutions are time-invariant, i.e., which supporting subspaces do not depend on time. Therefore, it is not surprising that in the literature these solutions have been associated with subspaces that, according to our terminology, would be denoted as “A-invariant supporting subspaces.” Analogously, in the study of the periodic solutions of the periodically time-varying Riccati equation, the key task consists in the classification of the supporting subspaces that are periodic with the same period as the coefficients of the equation.

Appendix. Herein, we report the statements of some technical lemmas that are needed throughout the paper. The proofs, which rely mostly on simple matrix manipulations, can be found in [33].

LEMMA A1. *Let $X(\cdot)$ be an SNS solution of the RDE (3). Then, $C(t)X(t)C(t)' + R(t) > 0$, for all t .*

LEMMA A2. *Let $R > 0$ and $\Pi \geq 0$ be such that $[R - C\Pi C']^{-1} > 0$ and let $P = \Pi + \Pi C'[R - C\Pi C']^{-1}C\Pi$. Then, $\Pi = P - PC'[R + CPC']^{-1}CP$.*

LEMMA A3. *Let $\Pi = P - PC'[R + CPC']^{-1}CP$, with $P \geq 0$ and $R > 0$. Then, $P = \Pi + \Pi C'[R - C\Pi C']^{-1}C\Pi$.*

LEMMA A4. *Let $\Gamma = I - PC'[R + CPC']^{-1}C$, with $R > 0$ and $P \geq 0$. Then, $\det \Gamma \neq 0$.*

LEMMA A5. *Assume that $Q(\cdot)$ is bounded and $A(\cdot)$ is exponentially stable. Let $X(\cdot)$ be an infinite-time SPS solution of the LDE*

$$X(t+1) = A(t)X(t)A(t)' + Q(t).$$

Then, if there exists a constant $k_1 > 0$, such that $X(t) \geq k_1 I$, for all t , the pair $(A(\cdot), Q(\cdot))$ is uniformly reachable.

Acknowledgments. The author acknowledges two anonymous reviewers for their helpful comments and suggestions.

REFERENCES

- [1] R. E. KALMAN, *New methods in Wiener filtering theory*, in Proc. 1st Sympos. on Engrg. Applications of Random Function Theory and Probability, J. Bogdanoff and F. Kozin, eds., John Wiley, New York, 1963, pp. 270–388.
- [2] R. S. BUCY, *Global theory of the Riccati equation*, J. Comput. Systems Sci., 1 (1967), pp. 349–361.
- [3] W. M. WONHAM, *On a matrix Riccati equation for stochastic control*, SIAM J. Control, 6 (1968), pp. 681–698.
- [4] V. KUCERA, *A contribution to matrix quadratic equations*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 344–347.
- [5] P. E. CAINES AND D. Q. MAYNE, *On the discrete-time matrix Riccati equation of optimal control*, Internat. J. Control, 12 (1970), pp. 785–794; 14 (1971), pp. 205–207.
- [6] J. C. WILLEMS, *Least-squares stationary optimal control and algebraic Riccati equation*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 621–634.
- [7] K. MARTENSSON, *On the matrix Riccati equation*, Inform. Sci., 3 (1971), pp. 17–49.
- [8] V. KUCERA, *On nonnegative definite solutions to matrix quadratic equations*, Automatica, 8 (1972), pp. 1413–1423.
- [9] B. P. MOLINARI, *The stabilizing solution of the algebraic Riccati equation*, SIAM J. Control, 11 (1973), pp. 262–271.
- [10] F. M. CALLIER AND J. L. WILLEMS, *Criterion for the convergence of the solution of the Riccati differential equation*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 1232–1242.
- [11] S. W. CHAN, G. C. GOODWIN, AND K. S. SIN, *Convergence properties of the Riccati difference equation in optimal filtering of nonstabilizable systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 110–118.
- [12] C. E. DE SOUZA, M. R. GEVERS, AND G. C. GOODWIN, *Riccati equations in optimal filtering of nonstabilizable systems having singular state transition matrices*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 831–838.
- [13] W. A. COPPEL, *Matrix quadratic equations*, Bull. Australian Math. Soc., 10 (1974), pp. 377–401.
- [14] M. A. SHAYMAN, *Geometry of the algebraic Riccati equation, Parts I and II*, SIAM J. Control Optim., 21 (1983), pp. 375–394, 395–409.
- [15] J. J. DEYST, JR. AND C. F. PRICE, *Conditions for asymptotic stability of the discrete minimum-variance linear estimator*, IEEE Trans. Automat. Control, AC-18 (1968), pp. 702–705; AC-18 (1973), pp. 562–563.
- [16] W. W. HAGER AND L. H. HOROWITZ, *Convergence and stability properties of the discrete Riccati operator equation and the associated optimal control and filtering problems*, SIAM J. Control Optim., 14 (1976), pp. 295–312.
- [17] 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.
- [18] G. DA PRATO AND A. ICHIKAWA, *Quadratic control for linear time-varying systems*, SIAM J. Control Optim., 28 (1990), pp. 359–381.
- [19] H. KANO AND T. NISHIMURA, *Periodic solutions of matrix Riccati equations with detectability and stabilizability*, Internat. J. Control, 29 (1979), pp. 471–487.
- [20] M. A. SHAYMAN, *On the phase portrait of the matrix Riccati equation arising from the periodic control problem*, SIAM J. Control Optim., 23 (1985), pp. 717–751.
- [21] C. E. DE SOUZA, *Riccati differential equation in optimal filtering of periodic non-stabilizable systems*, Internat. J. Control, 46 (1987), pp. 1235–1250.
- [22] 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.
- [23] J. C. ENGWERDA, *Stabilizability and detectability of discrete-time time-varying systems*, IEEE Trans. Automat. Control, AC-35 (1990), pp. 425–429.
- [24] B. D. O. ANDERSON AND J. B. MOORE, *Comments on “stabilizability and detectability of discrete-time time-varying systems,”* IEEE Trans. Automat. Control, to appear.
- [25] ———, *New results in linear system stability*, SIAM J. Control, 23 (1969), pp. 398–414.
- [26] D. L. KLEINMAN, *On an iterative technique for Riccati equation computations*, IEEE Trans. Automat. Control, AC-13 (1968), pp. 114–115.

- [27] N. R. SANDELL, JR., *On Newton's method for Riccati equation solution*, IEEE Trans. Automat. Control, AC-19 (1972), pp. 254–255.
- [28] G. A. HEWER, *An iterative technique for the computation of the steady state gains for the discrete optimal regulator*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 382–384.
- [29] R. S. BUCY AND P. D. JOSEPH, *Filtering for Stochastic Processes with Applications to Guidance*, John Wiley, New York, 1968.
- [30] R. R. BITMEAD, M. R. GEVERS, I. R. PETERSEN, AND R. J. KAYE, *Monotonicity and stabilizability properties of solutions of the Riccati difference equation: Propositions, lemmas, theorems, fallacious conjectures and counterexamples*, Systems Control Lett., 5 (1985), pp. 309–315.
- [31] T. NISHIMURA, *On the a priori information in sequential estimation problems*, IEEE Trans. Automat. Control, AC-11 (1966), pp. 197–204; AC-12 (1967), p. 123.
- [32] C. E. DE SOUZA, *On stabilizing properties of solutions of the Riccati difference equation*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 1313–1316.
- [33] G. DE NICOLAO, *On the time-varying Riccati difference equation of optimal filtering*, Internat. Rep. 90-02, Centro Teoria dei Sistemi—CNR, Dipartimento di Elettronica, Politecnico di Milano, Italy, 1990.

VERIFICATION OF THE SELF-STABILIZATION MECHANISM IN ROBUST STOCHASTIC ADAPTIVE CONTROL USING LYAPUNOV FUNCTION ARGUMENTS*

MILOJE RADENKOVIC†‡ AND ANTHONY N. MICHEL†

Abstract. The objective of this paper is to propose a new algorithm for self-tuning control in the presence of unmodeled dynamics. The algorithm is a modified version of the well-known stochastic gradient scheme. It is shown (with probability one) that the resulting closed-loop system is globally stable and the mean-square tracking error is proportional to the size of unmodeled dynamics. In the absence of unmodeled dynamics, the algorithm produces the minimum-variance self-tuning control. It is analytically verified that the proposed algorithm has self-stabilization property; i.e., possible occurrence of instability results in mean-square bounded signals. Global stability of the adaptive system is achieved without imposing persistency exciting condition on the regressor and positive real assumption on the system noise dynamics.

Key words. robust adaptive control, stochastic systems, self-stabilization

AMS(MOS) subject classifications. 93C40, 93E10

1. Introduction. It has long been recognized that the presence of unmodeled dynamics can cause a degradation of performance and instability of otherwise satisfactory adaptive control algorithms. This disturbing fact has been discovered in deterministic models [1], [2], and a large number of results have since been obtained to design a robust adaptive control in this context [3], [4]. The well-known ways of neutralizing the effects of unmodeled dynamics, such as the σ -modification, signal normalization, (relative) dead zone, and projection methods, have been widely used and discussed in the literature (see, for example, [4]). The intermittency phenomenon has also been encountered in practice in adaptive systems associated with unknown disturbances or unmodeled effects. Different kinds of unstable behavior and possible self-stabilization of the adaptive systems were first described in [1], [5]–[7], and later in [8]–[14].

In the stochastic environment, the most powerful adaptive control algorithms [15], [16] have been almost exclusively developed for the “ideal case” without unmodeled dynamics. This is not surprising, since none of the standard robustness measures, which worked so well for deterministic models, can be used in adaptive control of stochastic systems. The main reason is that the gain sequence of the algorithm must converge to zero even if the unmodeled dynamics is absent. The choice of the gain sequence is further complicated if the modeling errors are present because the speed of convergence should be low enough to capture the new information contained in the measurement vector and, at the same time, provide a normalization effect. These difficulties, however, can be avoided if the analysis is restricted at the outset to systems that are open-loop stable [17]. Recently, very interesting results have been presented in [18], where the authors propose robust control algorithms based on the signal normalization philosophy.

The main objective of this paper is to propose a new algorithm for stochastic adaptive control in the presence of unmodeled dynamics. The algorithm is of the self-tuning stochastic approximation variety, with two major modifications. First, a

* Received by the editors December 26, 1990; accepted for publication (in revised form) June 16, 1991. This work was supported by National Science Foundation grant ECS-88-02924.

† Department of Electrical Engineering, University of Notre Dame, Notre Dame, Indiana 46556.

‡ Present address, Department of Electrical Engineering, University of Colorado at Denver, 1200 Larimer Street, Denver, Colorado 80204-5300.

projection mechanism is used to bound the estimates of the controller parameters. Second, a sufficiently slow rate of the gain sequence is chosen without loss of the normalization effect. At the same time, the rate is fast enough to guarantee (asymptotic) optimality of the self-tuning control when the unmodeled dynamics disappear. It is shown that the proposed algorithm possesses a *self-stabilization property*. Since unmodeled dynamics tend to destabilize the system, over some operation periods the algorithm can exhibit an unstable behavior; i.e., regulator parameters may escape the stabilizing set, and, consequently, tracking errors start to grow without bounds. Subsequently, a decreasing Lyapunov function can be constructed, implying the stability of the tracking error. Specifically, bursts in the tracking error produce high-level and more exciting signals, thus forcing the estimator to estimate regulator parameters correctly. After the parameter estimates reenter the stability region, tracking error becomes stable in the mean-square sense. This self-stabilization effect is analytically evaluated without requiring the persistency excitation condition to be satisfied. Finally, repeating itself whenever instability occurs, the self-stabilization mechanism provides a globally stable closed-loop system.

The proposed algorithm ensures global stability of the closed-loop system even if the size of the unmodeled dynamics is large, perhaps at the price of high-gain feedback. Consequently, in the ideal case, the same type of stability is established without the standard strict positive realness (SPR) condition. This is achieved by forcing the SPR condition on the system and treating the rest of the noise as a mean-square bounded disturbance. We should note that global stability in this context has been established by choosing suitable stochastic Lyapunov functions and accommodating (by now standard) Goodwin's methodology [15], based upon the martingale convergence theory.

2. Problem statement. Let us consider the following stochastic, discrete-time, single-input, single-output (SISO) system with unmodeled dynamics:

$$(2.1) \quad A(q^{-1})y(t) = q^{-d}B(q^{-1})u(t) + C(q^{-1})\omega(t) + \gamma(t-1),$$

where $\{y(t)\}$, $\{u(t)\}$, and $\{\omega(t)\}$ are output, input, and stochastic disturbance sequences, respectively; q^{-1} represents the unit delay operator, while d is the pure time delay of the exactly modeled system part. Polynomials $A(q^{-1})$, $B(q^{-1})$, and $C(q^{-1})$ are given by

$$\begin{aligned} A(q^{-1}) &= 1 + a_1q^{-1} + \cdots + a_{n_A}q^{-n_A}, \\ B(q^{-1}) &= b_0 + b_1q^{-1} + \cdots + b_{n_B}q^{-n_B} \quad (b_0 \neq 0), \\ C(q^{-1}) &= 1 + c_1q^{-1} + \cdots + c_{n_C}q^{-n_C}. \end{aligned}$$

The unmodeled dynamics $\gamma(t)$ are assumed to be represented by

$$(2.2) \quad |\gamma(t)| \leq \gamma \sum_{j=1}^t \lambda_\gamma^{t-j} (|y(j)| + |\omega(j)| + k_\gamma), \quad \gamma > 0, \quad 0 < \lambda_\gamma < 1, \quad 0 \leq k_\gamma < \infty.$$

We adopt the following assumption concerning system (2.1):

$$(A_1) \quad B(z^{-1}) \text{ has zeros strictly outside the unit disc.}$$

Let $\omega(t)$ be a stochastic process defined on the underlying probability space $\{\Omega, \mathcal{F}, P\}$. We introduce the following assumptions:

$$(A_2) \quad \text{If } \mathcal{F}_t \text{ is the } \sigma\text{-algebra generated by } \{\omega(1), \dots, \omega(t)\}, \text{ then, for } t \geq 1,$$

$$E\{\omega(t+1) | \mathcal{F}_t\} = 0 \quad (\text{a.s.}),$$

$$E\{\omega(t+1)^2 | \mathcal{F}_t\} = \sigma_\omega^2 \quad (\text{a.s.}),$$

$$\sup_t E\{|\omega(t+1)|^{2+\eta} | \mathcal{F}_t\} \leq k_\omega < \infty, \quad \eta > 0 \quad (\text{a.s.}).$$

Our objective is to design a controller as a function of initial conditions and measurements to stabilize the system and, for a given reference signal $y^*(t)$, to minimize the functional criterion

$$(2.3) \quad J = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=1}^N (y(t) - y^*(t))^2$$

without requiring explicit knowledge of the system model. We assume that reference signal $y^*(t)$ satisfies

$$(A_3) \quad \{y^*(t)\} \text{ is a bounded deterministic sequence defined for } t \geq 1; \text{ that is, there exists a number } m_1 \text{ such that } |y^*(t)| \leq m_1 \text{ for all } t \geq 1.$$

Note that, for a given $n_C = \deg C(q^{-1})$, there exists a polynomial

$$(2.4) \quad \tilde{C}(q^{-1}) = 1 + \tilde{c}_1 q^{-1} + \dots + \tilde{c}_{n_C} q^{-n_C}$$

of the same degree as polynomial $C(q^{-1})$, so that, for some $0 < \bar{a} < 1$, $\tilde{C}(z^{-1}) - \bar{a}/2$ is a strictly positive real function. If polynomials $F(q^{-1})$ and $G(q^{-1})$ are the minimum degree solutions with respect to $F(q^{-1})$, of the diophantine equation

$$(2.5) \quad \tilde{C}(q^{-1}) = A(q^{-1})F(q^{-1}) + q^{-d}G(q^{-1})$$

and

$$(2.6) \quad \deg F(q^{-1}) = n_F \leq d-1 \text{ and } \deg G(q^{-1}) = n_G \leq \max\{n_A - 1; n_C - d\},$$

then, from (2.1), we obtain that

$$(2.7) \quad \begin{aligned} &\tilde{C}(q^{-1})\{y(t+d) - y^*(t+d) - F(q^{-1})\omega(t+d)\} \\ &= B(q^{-1})F(q^{-1})u(t) + G(q^{-1})y(t) - \tilde{C}(q^{-1})y^*(t+d) + \nu(t+d-1), \end{aligned}$$

where

$$(2.8) \quad \nu(t-1) = F(q^{-1})\{\gamma(t-1) + [C(q^{-1}) - \tilde{C}(q^{-1})]\omega(t)\}.$$

In (2.8) we actually separate from $C(z^{-1})$ its strict positive real part $\tilde{C}(z^{-1})$. The remaining part $C(z^{-1}) - \tilde{C}(z^{-1})$ produces this term $(C(q^{-1}) - \tilde{C}(q^{-1}))\omega(t)$, which acts as an external mean-square bounded disturbance. Note that we are not assuming a *positive real or stability condition* on the $C(z^{-1})$.

If the system is completely modeled, i.e., when $\gamma(t) \equiv 0$ for $t \geq 0$ and if $\tilde{C}(q^{-1}) = C(q^{-1})$ from (2.8), it follows that the controller optimal in the sense of (2.3) is given by

$$(2.9) \quad B(q^{-1})F(q^{-1})u(t) = -G(q^{-1})y(t) + \tilde{C}(q^{-1})y^*(t+d)$$

and the achieved minimal value of the criterion (2.3) is $J_{\min} = \sigma_1^2$, where

$$(2.10) \quad \sigma_1^2 = E\{[F(q^{-1})\omega(t+d)]^2 | \mathcal{F}_t\}.$$

3. Robust adaptive control. We assume that parameters of the exactly modeled system part, as well as unmodeled dynamics in (2.1), are not known to the designer,

i.e., that $u(t)$ can only implicitly depend on them through observations. Let us define the parameters of the controller as

$$(3.1) \quad \theta_0^T = \begin{bmatrix} g_0, \dots, g_{n_G}, 0, \dots, 0; \\ n_1 - n_G, \\ r_0, r_1, \dots, r_{n_R}, 0, \dots, 0; \\ n_2 - n_B - n_F, \tilde{c}_1, \dots, \tilde{c}_{n_3} \end{bmatrix},$$

where $n_1 = \max \{n_A - 1; n_C - d\}$, $n_2 = n_B + d - 1$, and $n_3 = n_C$, while g_i and r_i are the coefficients of the polynomials $G(q^{-1})$ and $R(q^{-1}) = B(q^{-1})F(q^{-1})$, respectively, with $n_R = n_B + n_F$. The definition of θ_0 allows us to rewrite (2.7) as

$$(3.2) \quad \tilde{C}(q^{-1})z(t) = \theta_0^T \phi(t) - y^*(t+d) + \nu(t+d-1),$$

where

$$(3.3) \quad \begin{aligned} z(t) &= y(t+d) - y^*(t+d) - F(q^{-1})\omega(t+d), \\ \phi(t)^T &= [y(t), \dots, y(t-n_1); u(t), \dots, u(t-n_2); \\ &\quad -y^*(t+d-1), \dots, -y^*(t+d-n_3)]. \end{aligned}$$

Observe that the control law (2.9) is equivalent to

$$(3.4) \quad \theta_0^T \phi(t) = y^*(t+d).$$

To construct an adaptive control algorithm, let us introduce the next assumption:

- (A₄) The compact convex set Θ^0 that contains the true parameters θ_0 , the sign of b_0 , and a lower bound $b_{0,\min}$ on the magnitude of b_0 are known.

For the estimation of θ_0 , we propose the following stochastic gradient-type algorithm:

$$(3.5) \quad \hat{\theta}(t+d) = F \left\{ \hat{\theta}(t) + \frac{\bar{a}}{\tilde{r}(t)} \phi(t) [y(t+d) - y^*(t+d)] \right\}, \quad 0 < \bar{a} < 1,$$

where $F\{\cdot\}$ projects orthogonally onto Θ^0 , so that $F\{\theta\} \in \Theta^0$ for all $\theta \in R^{n_1+n_2+n_3+2}$, and there exists a finite constant d_0 , so that $\|\hat{\theta}(t) - \theta_0\|^2 \leq d_0$ and $|\hat{b}_0(t)| \geq b_{0,\min} > 0$ for all $t > 0$. The algorithm gain sequence $\tilde{r}(t)$ is given by

$$(3.6) \quad \tilde{r}(t) = \max \left\{ 2 \max_{1 \leq \tau \leq t} \|\phi(\tau)\|^2, r(t)^{1-\varepsilon} + t^{1-\varepsilon} \right\}, \quad 0 < \varepsilon < \frac{1}{3},$$

where

$$(3.7) \quad r(t) = r(t-1) + \|\phi(t)\|^2, \quad r(0) > 1.$$

Since θ_0 is unknown, as adaptive control law we use the ‘‘certainty equivalence’’ version of (3.4), i.e.,

$$(3.8) \quad \hat{\theta}(t)^T \phi(t) = y^*(t+d).$$

Obviously, the estimation algorithm consists of d interlaced stochastic gradient-type procedures with (almost sure) finite initial conditions

$$\hat{\theta}(k) = \hat{\theta}_k, \quad k = 0, 1, \dots, d-1.$$

Let us provide a motivation for our choice of the adaptive control algorithm. It differs from the standard stochastic approximation algorithm [15] in two aspects. First, we introduce a modification $1/\tilde{r}(t)$ of the usual gain sequence $1/r(t)$, which converges to

zero more slowly than $1/r(t)$ and, at the same time, achieves the normalization of the regressor $\phi(t)$. By slowing down the gain sequence, we allow for the present measurements to have larger emphasis on the parameter update, thus weakening the effects of unmodeled dynamics. Intuitively, the proposed algorithm starts as a gradient normalization algorithm to become a type of stochastic approximation algorithm with increasing time.

The second aspect of the algorithm is the projection mechanism $F\{\cdot\}$, which prevents divergence of the parameter estimates caused by unmodeled dynamics and external disturbances.

Let us define positive constants

$$(3.9) \quad \begin{aligned} C_{BA} &= \max_{|z|=1} \left| \frac{A(z)}{B(z)} \right|, \quad C_B = \max_{|z|=1} \left| \frac{1}{B(z)} \right|, \quad C_{BC} = \max_{|z|=1} \left| \frac{C(z)}{B(z)} \right|, \\ C_F &= \max_{|z|=1} |F(z)|, \quad C_\Delta = \max_{|z|=1} |C(z) - \tilde{C}(z)|. \end{aligned}$$

Since the corresponding operators are stable, all constants defined by (3.9) are finite.

Henceforth, by ξ_i , $i = 0, 1, \dots$, $0 < \xi_i < \infty$, we will denote the effect of the initial conditions.

The following lemma will be useful for future reference.

LEMMA 3.1. (1) *It holds that*

$$(3.10) \quad z(t)^2 \leq C_\theta \max_{1 \leq \tau \leq t} z(\tau-1)^2 + l_0(t), \quad 0 < C_\theta < \infty,$$

where $z(t)$ is defined by (3.3), and

$$(3.11) \quad l_0(t) = k'_\theta \max_{1 \leq \tau \leq t} \omega(\tau+d)^2 + k''_\theta, \quad 0 < k'_\theta, \quad k''_\theta < \infty.$$

(2)(i) *The inequality*

$$(3.12) \quad |\gamma(t)| \leq \gamma_z(t) + \gamma_\omega(t)$$

holds, where

$$(3.13) \quad \gamma_z(t) = \gamma \sum_{j=1}^t \lambda_\gamma^{t-j} |z(j-d)|$$

and

$$(3.14) \quad \gamma_\omega(t) = \gamma \sum_{j=1}^t \lambda_\gamma^{t-j} (|y^*(j) + F(q^{-1})\omega(j)| + |\omega(j)| + \kappa_\gamma).$$

(ii) *The inequality*

$$(3.15) \quad |\nu(t)| \leq \nu_z(t) + \nu_\omega(t)$$

holds, where

$$(3.16) \quad \nu_z(t) = \sum_{i=0}^{n_F} |f_i| \gamma_z(t-i), \quad f_0 \equiv 1$$

and

$$(3.17) \quad \nu_\omega(t) = \sum_{i=0}^{n_F} |f_i| \gamma_\omega(t-i) + |F(q^{-1})[C(q^{-1}) - \tilde{C}(q^{-1})]\omega(t)|,$$

while $f_i, i = 0, 1, \dots, n_F$ are the coefficients of the polynomial $F(q^{-1})$, which is given by the polynomial equation (2.5).

(iii) The inequality

$$(3.18) \quad \sum_{t=1}^N |z(t)| \cdot |\nu_z(t+d-1)| \leq C_\gamma \sum_{t=1}^N z(t)^2 + \xi_0$$

holds, where

$$(3.19) \quad C_\gamma = \frac{\gamma C_f}{1 - \lambda_\gamma} \quad \text{and} \quad C_f^2 = (n_F + 1) \sum_{i=0}^{n_F} f_i^2.$$

(3) Finally,

$$(3.20) \quad \sum_{t=1}^N \|\phi(t)\|^2 \leq K_1 \sum_{t=1}^N z(t)^2 + K_2 N, \quad 0 < K_2 < \infty, \quad (\text{a.s.}),$$

where

$$(3.21) \quad K_1 = 2 \left\{ n_1 + 1 + 3(n_2 + 1) \left(C_{BA}^2 + \frac{3C_B^2 \gamma^2}{(1 - \lambda_\gamma)^2} \right) \right\}.$$

Proof. The proof of the lemma is given in the Appendix.

Concerning the size of unmodeled dynamics, we need the following assumption:

(A₅) $\rho_1 = \rho_m - C_\gamma > 0$, where C_γ is defined by (3.19), and ρ_m is the largest number so that $\tilde{C}(z^{-1}) - \bar{a}/2 - \rho_m$ remains a positive real function.

Observe that the definition of ρ_m implies that

$$(3.22) \quad S(t+d) = 2\bar{a} \sum_{j=1}^t z(j) \left\{ \tilde{C}(q^{-1}) - \frac{\bar{a}}{2} - \rho_m \right\} z(j) + K_3 \geq 0$$

for all $t \geq 0$ and for some $0 \leq K_3 < \infty$.

In this section, we prove the global stability of the proposed adaptive algorithm and evaluate the mean-square tracking error.

Let us define a sequence

$$(3.23) \quad W_0(t+d) = \bar{a} \sum_{j=1}^t \{ \rho_1 z(j)^2 - 2|z(j)| \cdot |\nu_\omega(j+d-1)| - 2C_1[r(j)^{1-\varepsilon} - r(j-1)^{1-\varepsilon}] \},$$

where ε is defined in (3.6), and $0 < C_1 < \infty$.

Essential to the convergence of the algorithm is the behavior of the sequence $W_0(t)$. First, we show that if $W_0(t) > 0$ for all $t \geq 1$, a suitably constructed Lyapunov function decreases, and, consequently, the algorithm is globally stable. If $W_0(t) \leq 0$ for all $t \geq 1$, boundedness of the mean-square tracking error by some quantity $\mathcal{O}(\gamma^2, C_\Delta)$ can be derived trivially from the definition of $W_0(t)$. Our attention will be concentrated on the case when $W_0(t)$ changes sign. In the intervals where $W(t) \leq 0$, we can easily conclude that mean-square tracking error is of order (γ^2, C_Δ) . Since, in the presence of unmodeled dynamics and external disturbances, the algorithm has a tendency to diverge, the tracking error becomes large enough so that $W_0(t)$ becomes positive. Then, as we will see, there exists a decreasing Lyapunov function, implying l_2 boundedness of the tracking error.

For the purpose of our analysis, the following lemma will also prove to be useful.

LEMMA 3.2. *Let assumptions (A₁)–(A₅) hold. Then, on the subsequence $\{N_k\}$, $k = 1, 2, 3, \dots$, where $W_0(N_k + d) > 0$,*

$$(3.24) \quad \begin{aligned} S(N_k + d) + \bar{a}\rho_1 \sum_{t=1}^{N_k} z(t)^2 \\ \leq d_0 d\tilde{r}(N_k) + o[(r(N_k) + N_k)^{1-\varepsilon}] + K_4, \quad (a.s.), \end{aligned}$$

where $0 < K_4 < \infty$, and ρ_1 is defined by assumption (A₅).

Proof. Starting from (3.5), we can obtain that

$$(3.25) \quad V(t+d) \leq V(t) + \frac{2\bar{a}}{\tilde{r}(t)} \tilde{\theta}(t)^T \phi(t) e(t+d) + \frac{\bar{a}^2 \|\phi(t)\|^2 e(t+d)^2}{\tilde{r}(t)^2},$$

where

$$(3.26) \quad V(t) = \tilde{\theta}(t)^T \tilde{\theta}(t) \quad \text{and} \quad \hat{\theta}(t) = \tilde{\theta}(t) - \theta_0, \quad e(t) = y(t) - y^*(t).$$

Using the definition of $z(t)$, from (3.25) it is obvious that

$$(3.27) \quad \begin{aligned} V(t+d) \leq V(t) + \frac{2\bar{a}}{\tilde{r}(t)} \tilde{\theta}(t)^T \phi(t) z(t) + \frac{2\bar{a}}{\tilde{r}(t)} \tilde{\theta}(t)^T \phi(t) F(q^{-1}) \omega(t+d) \\ + \frac{2\bar{a}^2 \|\phi(t)\|^2}{\tilde{r}(t)^2} \{z(t)^2 + [F(q^{-1}) \omega(t+d)]^2\}. \end{aligned}$$

Note that, from (3.2) and (3.8),

$$(3.28) \quad \tilde{C}(q^{-1})z(t) = -\tilde{\theta}(t)^T \phi(t) + \nu(t+d-1),$$

where $\nu(t)$ is defined by (2.8). Since $2\|\phi(t)\|^2/\tilde{r}(t) \leq 1$, by combining (3.27) and (3.28) after simple majorizations, we obtain that

$$(3.29) \quad \begin{aligned} V(t+d) \leq V(t) - \frac{2\bar{a}}{\tilde{r}(t)} \left\{ \left(C(q^{-1}) - \frac{\bar{a}}{2} \right) z(t) \right\} z(t) \\ + \frac{1}{\tilde{r}(t)} 2\bar{a}z(t)\nu(t+d-1) \\ + \frac{2\bar{a}}{\tilde{r}(t)} \tilde{\theta}(t)^T \phi(t) F(q^{-1}) \omega(t+d) \\ + \frac{2\bar{a}^2 \|\phi(t)\|^2}{\tilde{r}(t)^2} [F(q^{-1}) \omega(t+d)]^2. \end{aligned}$$

Considering the definition of $S(t)$ (3.22), from (3.29) we obtain that

$$(3.30) \quad \begin{aligned} V(t+d) + \frac{S(t+d)}{\tilde{r}(t)} \leq V(t) + \frac{S(t+d-1)}{\tilde{r}(t)} - 2\bar{a}\rho_m \frac{z(t)^2}{\tilde{r}(t)} \\ + \frac{2\bar{a}z(t)\nu(t+d-1)}{\tilde{r}(t)} \\ + \frac{2\bar{a}}{\tilde{r}(t)} \tilde{\theta}(t)^T \phi(t) F(q^{-1}) \omega(t+d) \\ + \frac{2\bar{a}^2 \|\phi(t)\|^2}{\tilde{r}(t)^2} [F(q^{-1}) \omega(t+d)]^2. \end{aligned}$$

From the previous relation, by using statement 2(ii) of Lemma 3.1, we obtain that

$$\begin{aligned}
 V(t+d)\tilde{r}(t) + S(t+d) &\leq V(t)\tilde{r}(t-d) + d_0(\tilde{r}(t) - \tilde{r}(t-d)) \\
 &\quad + S(t+d-1) - 2\bar{a}\rho_m z(t)^2 + 2\bar{a}|z(t)| \cdot |\nu_z(t+d-1)| \\
 (3.31) \quad &\quad + 2\bar{a}|z(t)| \cdot |\nu_\omega(t+d-1)| + 2\bar{a}\tilde{\theta}(t)^T \phi(t) F(q^{-1})\omega(t+d) \\
 &\quad + \frac{2\bar{a}^2 \|\phi(t)\|^2}{\tilde{r}(t)} [F(q^{-1})\omega(t+d)]^2,
 \end{aligned}$$

where we used the fact that $V(t) \leq d_0$ and $\tilde{r}(t) \geq \tilde{r}(t-1)$. After summation from $t=1$ to N , by using statement 2(iii) of Lemma 3.1, we obtain that

$$\begin{aligned}
 V(N+d)\tilde{r}(N) + \dots + V(N+1)\tilde{r}(N-d+1) + S(N+d) \\
 \leq V(1)\tilde{r}(1-d) + \dots + V(d)\tilde{r}(1) + S(d) + d_0 d\tilde{r}(N) \\
 (3.32) \quad - 2\bar{a}(\rho_m - C_\gamma) \sum_{t=1}^N z(t)^2 + 2\bar{a} \sum_{t=1}^N |z(t)| \cdot |\nu_\omega(t+d-1)| \\
 + 2\bar{a} \sum_{t=1}^N \tilde{\theta}(t)^T \phi(t) F(q^{-1})\omega(t+d) \\
 + 2\bar{a}^2 \sum_{t=1}^N \frac{\|\phi(t)\|^2}{\tilde{r}(t)} [F(q^{-1})\omega(t+d)]^2.
 \end{aligned}$$

Observe that, by statement (2) of the lemma in the Appendix,

$$(3.33) \quad \sum_{t=1}^N \frac{\|\phi(t)\|^2}{\tilde{r}(t)} [F(q^{-1})\omega(t+d)]^2 \leq o[(r(N) + N)^{2\epsilon}] \quad (\text{a.s.}).$$

Since $\tilde{\theta}(t)^T \phi(t)$ is \mathcal{F}_t measurable, the local martingale convergence theorem (LMCT) yields

$$(3.34) \quad \left| \sum_{j=1}^t \tilde{\theta}(j)^T \phi(j) F(q^{-1})\omega(j+d) \right| \leq C_1 r(t)^{1-\epsilon} \quad (\text{a.s.}),$$

where C_1 is the same constant as in (3.23) and ϵ is defined by (3.6). Note that LMCT is valid for all $\frac{1}{2} < 1-\epsilon < 1$, and ϵ from (3.6) satisfies this condition. Using the least two relations, from (3.32), we obtain that

$$\begin{aligned}
 S(N+d) + 2\bar{a}\rho_1 \sum_{t=1}^N z(t)^2 - 2\bar{a} \sum_{t=1}^N |z(t)| \cdot |\nu_\omega(t+d-1)| \\
 (3.35) \quad \leq V(1)\tilde{r}(1-d) + \dots + V(d)\tilde{r}(1) + S(d) \\
 + d_0 d\tilde{r}(N) + 2\bar{a}C_1 r(N)^{1-\epsilon} + o[(r(N) + N)^{2\epsilon}] \quad (\text{a.s.}).
 \end{aligned}$$

On the subsequence $\{N_k\}$, where $W_0(N_k + d) > 0$, from the definition of $W_0(N+d)$, we derive

$$\begin{aligned}
 (3.36) \quad \bar{a} \sum_{t=1}^{N_k} \{\rho_1 z(t)^2 - 2|z(t)| \cdot |\nu_\omega(t+d-1)|\} - 2\bar{a}C_1 r(N_k)^{1-\epsilon} \\
 \geq -2\bar{a}C_1 r(0)^{1-\epsilon},
 \end{aligned}$$

where ρ_1 is defined by assumption (A₅).

Thus, from (3.35) and (3.36), it follows that

$$(3.37) \quad S(N_k + d) + \bar{a}\rho_1 \sum_{t=1}^{N_k} z(t)^2 \leq d_0 d\tilde{r}(N_k) + o[(r(N_k) + N_k)^{2\varepsilon}] + K_4,$$

where $0 < K_4 < \infty$. Thus Lemma 3.2 is proved. \square

Global stability results will be formulated in the following theorem.

THEOREM 3.1. *Let assumptions (A_1) – (A_5) hold. Then*

$$(3.38) \quad (1) \quad \lim_{N \rightarrow \infty} \sup_N \frac{1}{N} \sum_{t=1}^N z(t)^2 \leq (1 + C_2) \Sigma_0, \quad 0 < C_2 < \infty,$$

where $z(t)$ is defined by (3.3) and

$$(3.39) \quad \Sigma_0 = \frac{16}{\rho_1^2} \{ \Sigma_\gamma + \Sigma_\Delta + 2(\Sigma_\gamma \Sigma_\Delta)^{1/2} \},$$

$$(3.40) \quad \Sigma_\gamma = 2C_\gamma^2(m_1 + \sigma_1 + \sigma_\omega + \kappa_\gamma)^2,$$

$$(3.41) \quad \Sigma_\Delta = C_F^2 C_\Delta^2 \sigma_\omega^2,$$

where ρ_1 is defined by assumption (A_5) , and C_γ is given by (3.19), while C_F and C_Δ are defined by (3.9). The following also hold:

$$(3.42) \quad (2) \quad \lim_{N \rightarrow \infty} \sup_N \frac{1}{N} \sum_{t=1}^N (y(t) - y^*(t))^2 \leq \sigma_1^2 + 2 \Sigma_0^{1/2} \sigma_1 + \Sigma_0 \quad (\text{a.s.}),$$

$$(3.43) \quad (3) \quad \lim_{N \rightarrow \infty} \sup_N \frac{1}{N} \sum_{t=1}^N \|\phi(t)\|^2 \leq C'_1 < \infty \quad (\text{a.s.}).$$

Proof. Observe that relation (3.34) implies that

$$(3.44) \quad W_1(t+d) = 2\bar{a} \left\{ C_1 r(t)^{1-\varepsilon} - \sum_{j=1}^t \tilde{\theta}(j)^T \phi(j) F(q^{-1}) \omega(j+d) \right\} \geq 0 \quad (\text{a.s.}).$$

Thus, from (3.30) by (3.15) and (3.44), we can get that

$$(3.45) \quad \begin{aligned} & V(t+d) + \frac{S(t+d) + W_1(t+d)}{\tilde{r}(t)} \\ & \leq V(t) + \frac{S(t+d-1) + W_1(t+d-1)}{\tilde{r}(t-1)} + \frac{2\bar{a}C_1[r(t)^{1-\varepsilon} - r(t-1)^{1-\varepsilon}]}{\tilde{r}(t)} \\ & \quad + \frac{2\bar{a}|z(t)| \cdot |v_z(t+d-1)|}{\tilde{r}(t)} + \frac{2\bar{a}|z(t)| \cdot |v_\omega(t+d-1)|}{\tilde{r}(t)} \\ & \quad - 2\bar{a}\rho_m \frac{z(t)^2}{\tilde{r}(t)} + \frac{2\bar{a}^2 \|\phi(t)\|^2}{\tilde{r}(t)^2} [F(q^{-1})\omega(t+d)]^2. \end{aligned}$$

From statement 2(iii) of Lemma 3.1, we derive, for all $t \geq 1$,

$$(3.46) \quad W_2(t+d) = 2\bar{a} \sum_{j=1}^t \{ C_\gamma z(j)^2 - |z(j)| \cdot |v_z(d-1)| \} + 2a\xi_0 \geq 0,$$

where C_γ is defined by (3.19).

Relations (3.45) and (3.46) imply that

$$\begin{aligned}
 & V(t+d) + \frac{S(t+d) + W_1(t+d) + W_2(t+d)}{\tilde{r}(t)} \\
 & \leq V(t) + \frac{S(t+d-1) + W_1(t+d-1) + W_2(t+d-1)}{\tilde{r}(t-1)} \\
 & \quad - 2\bar{a}\rho_1 \frac{z(t)^2}{\tilde{r}(t)} + \frac{2\bar{a}|z(t)| \cdot |\nu_\omega(t+d-1)|}{\tilde{r}(t)} \\
 & \quad + \frac{2\bar{a}C_1(r(t)^{1-\varepsilon} - r(t-1)^{1-\varepsilon})}{\tilde{r}(t)} \\
 & \quad + \frac{2\bar{a}^2 \|\phi(t)\|^2}{\tilde{r}(t)^2} [F(q^{-1})\omega(t+d)]^2,
 \end{aligned} \tag{3.47}$$

where ρ_1 is defined by assumption (A₅).

Let us first consider the case where $W_0(t+d) > 0$ for all $t \geq 0$, where $W_0(t)$ is defined by (3.23). Then, from (3.47), we obtain that

$$\begin{aligned}
 & V(t+d) + \frac{S(t+d) + W_0(t+d) + W_1(t+d) + W_2(t+d)}{\tilde{r}(t)} \\
 & \leq V(t) + \frac{S(t+d-1) + W_0(t+d-1) + W_1(t+d-1) + W_2(t+d-1)}{\tilde{r}(t-1)} \\
 & \quad - \bar{a}\rho_1 \frac{z(t)^2}{\tilde{r}(t)} + \frac{2\bar{a} \|\phi(t)\|^2}{\tilde{r}(t)^2} [F(q^{-1})\omega(t+d)]^2.
 \end{aligned} \tag{3.48}$$

From the above relation, by the lemma in the Appendix, we obtain that

$$\lim_{N \rightarrow \infty} \sum_{t=1}^N \bar{a}\rho_1 \frac{z(t)^2}{\tilde{r}(t)} \leq C_3 < \infty \quad (\text{a.s.}), \tag{3.49}$$

wherefrom, by Kronecker's Lemma and (3.20), we get that

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=1}^N z(t)^2 = 0 \quad (\text{a.s.}); \tag{3.50}$$

i.e., in the case where $W_0(t) > 0$ for all t , the statements of the theorem are valid. Let us consider the case where $W_0(t+d) \leq 0$ for all $t \geq 0$. From the definition of $W_0(t)$, by using statement (3) of Lemma 3.1, we can conclude that

$$\begin{aligned}
 & \rho_1 \sum_{t=1}^N z(t)^2 \leq 4 \max \left\{ \sum_{t=1}^N |z(t)| \cdot |\nu_\omega(t+d-1)|; \right. \\
 & \quad \left. C_4 \left(\sum_{t=1}^N z(t)^2 \right)^{1-\varepsilon} + C_5 N^{1-\varepsilon} \right\} \quad (\text{a.s.}),
 \end{aligned} \tag{3.51}$$

where $0 < C_4, C_5 < \infty$.

If $\lim_{N \rightarrow \infty} \sum_{t=1}^N z(t)^2 \leq K_4 < \infty$, the statements of the theorem are true. If $\lim_{N \rightarrow \infty} \sum_{t=1}^N z(t)^2 = \infty$, then, for $0 < \rho_0 \ll \bar{a}\rho_1$,

$$\rho_0 \sum_{t=1}^N z(t)^2 \geq \mathcal{O} \left(\sum_{t=1}^N z(t)^2 \right)^\alpha, \quad 0 < \alpha < 1, \tag{3.52}$$

wherefrom it follows that

$$(3.53) \quad \begin{aligned} \sum_{t=1}^N z(t)^2 &\leq \max \left\{ \frac{16}{\rho_1^2} \sum_{t=1}^N \nu_\omega(t+d-1)^2, \frac{C_5 N^{1-\varepsilon}}{\bar{a}\rho_1 - \rho_0} \right\} \\ &\leq \frac{16}{\rho_1^2} \sum_{t=1}^N \nu_\omega(t+d-1)^2 \quad (\text{a.s.}), \end{aligned}$$

where we used the fact that, by assumption (A_2) ,

$$\sum_{t=1}^N \nu_\omega(t)^2 = \mathcal{O}(N) \quad (\text{a.s.}).$$

From (3.53), it follows that

$$(3.54) \quad \limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{t=1}^N z(t)^2 \leq \Sigma_0 \quad (\text{a.s.}),$$

where Σ_0 is defined by (3.39). All statements of the theorem directly follow from the previous relation.

Our major interest will be concentrated on the case when there are intervals where $W_0(t)$ is positive and intervals where $W_0(t)$ is nonpositive. Let us define the sequences τ_k and σ_k , $k = 1, 2, 3, \dots$, as follows:

$$1 \triangleq \tau_1 < \sigma_1 < \tau_2 < \sigma_2 < \dots < \tau_k < \sigma_k < \tau_{k+1} < \dots,$$

so that

$$(3.55) \quad W_0(t+d) > 0 \text{ for all } t \in [\sigma_k, \tau_{k+1}) \text{ and } W_0(t+d) \leq 0 \text{ for all } t \in [\tau_k, \sigma_k).$$

If $W_0(1+d) > 0$, then we set $\tau_1 = 0$ and $\sigma_1 = 1$. The following three cases are possible:

- (1) There exists some $k_0 < \infty$ so that $\tau_{k_0} < \infty$ and $\sigma_{k_0} = +\infty$;
- (2) There exists some $k_0 < \infty$ so that $\sigma_{k_0} < \infty$ and $\tau_{k_0+1} = +\infty$; and
- (3) $\tau_k < \infty$ and $\sigma_k < \infty$ for all finite k .

In the first case, we have from (3.35) that $W_0(t+1) \leq 0$ for $t \geq \tau_{k_0}$, and, in the second case, we have that $W_0(t+1) > 0$ for $t \geq \sigma_{k_0}$. These two cases are covered by the previous analysis. Next, we consider the case when $\tau_k < \infty$ and $\sigma_k < \infty$, for all finite k .

Since for $t \in [\sigma_k, \tau_{k+1})$, $W_0(t+d) > 0$ after summation from $t = \sigma_k + 1$ to $N < \tau_{k+1}$, we can obtain from (3.48) that

$$(3.56) \quad \begin{aligned} &V(N+d) + \dots + V(N+1) \\ &+ \frac{S(N+d) + W_0(N+d) + W_1(N+d) + W_2(N+d)}{\tilde{r}(N)} \\ &\leq V(\sigma_k+1) + \dots + V(\sigma_k+d) \\ &+ \frac{S(\sigma_k+d) + W_0(\sigma_k+d) + W_1(\sigma_k+d) + W_2(\sigma_k+d)}{\tilde{r}(\sigma_k)} \\ &- \bar{a}\rho_1 \sum_{t=\sigma_k+1}^N \frac{z(t)^2}{\tilde{r}(t)} + 2\bar{a}^2 \sum_{t=\sigma_k+1}^N \frac{\|\phi(t)\|^2}{\tilde{r}(t)^2} [F(q^{-1})\omega(t+d)]^2. \end{aligned}$$

Note that Lemma 3.2 is valid for all $N \in [\sigma_k, \tau_{k+1})$ and, consequently,

$$(3.57) \quad \frac{S(\sigma_k+d)}{\tilde{r}(\sigma_k)} \leq C_6 < \infty \quad (\text{a.s.}).$$

Since

$$(3.58) \quad W_0(\sigma_k + d) \leq \bar{a}\rho_1 \sum_{j=1}^{\sigma_k} z(j)^2$$

and

$$(3.59) \quad W_2(\sigma_k + d) \leq 2\bar{a}C_\gamma \sum_{j=1}^{\sigma_k} z(j)^2,$$

from Lemma 3.2, we derive

$$(3.60) \quad \frac{W_0(\sigma_k + d) + W_2(\sigma_k + d)}{\tilde{r}(\sigma_k)} \leq C_7 < \infty \quad (\text{a.s.}).$$

Observe also that by relation (3.34)

$$(3.61) \quad \frac{W_1(\sigma_k + d)}{\tilde{r}(\sigma_k)} \leq C_8 < \infty \quad (\text{a.s.}).$$

From relation (3.56), by (3.57), (3.60), (3.61), and statement (1) of the lemma in the Appendix, we get for $N \in [\sigma_k, \tau_{k+1})$ that

$$(3.62) \quad \bar{a}\rho_1 \sum_{t=\sigma_k+1}^N \frac{z(t)^2}{\tilde{r}(t)} \leq C_9 < \infty \quad (\text{a.s.}).$$

Let us first consider the subsequence $\{p_n\}$, $n = 1, 2, 3, \dots$, where $\lim_{n \rightarrow \infty} \sup(\tau_{p_n+1} - \sigma_{p_n}) \leq L_{1p} < \infty$. Since for $t \in [\tau_{p_n}, \sigma_{p_n})$ relation (3.55) is valid, similarly as in (3.53), we can get that

$$(3.63) \quad \sum_{j=1}^{\sigma_{p_n}-1} z(j)^2 \leq \frac{16}{\rho_1^2} \sum_{j=1}^{\sigma_{p_n}-1} \nu_\omega(j+d-1)^2 \quad (\text{a.s.}),$$

wherefrom it follows that

$$(3.64) \quad \max_{1 \leq \tau \leq \sigma_{p_n}-1} z(\tau)^2 \leq \alpha(\sigma_{p_n}-1) \quad (\text{a.s.}),$$

where

$$(3.65) \quad \alpha(\sigma_{p_n}-1) = \frac{16}{\rho_1^2} \sum_{j=1}^{\sigma_{p_n}-1} \nu_\omega(j+d-1)^2.$$

Using the “exponentially growing rule” (statement (1) of Lemma 3.1) and (3.64), we obtain, for $t < \tau_{k_n+1}$, that

$$(3.66) \quad \max_{1 \leq \tau \leq t} z(\tau)^2 \leq C_{\theta}^{L_{1p}} \alpha(\sigma_{p_n}-1) + \sum_{i=0}^{L_{1p}-1} C_{\theta}^i I_0(t-i) \quad (\text{a.s.}).$$

Let us now get a similar bound for $z(t)^2$, on the subsequence $\{k_m\}$, where $\lim_{m \rightarrow \infty} (\tau_{k_m+1} - \sigma_{k_m}) = +\infty$. Observe that, in this case, statement (3.62) implies that, for arbitrarily small $\varepsilon_2 > 0$,

$$(3.67) \quad z(t)^2 \leq \varepsilon_2 \tilde{r}(t) \quad (\text{a.s.}),$$

for $t \in [\sigma_{k_m}, \tau_{k_m+1})$, except for a finite number of points $t_1 < t_2 < \dots < t_L$, where $L = [C_9/\bar{a}\rho_1\varepsilon_2]$, $t_1 \leq \sigma_{k_m}$, and $t_L < \tau_{k_m+1}$. Denote by \mathcal{C}_L the set containing the points t_i , $i = 1, 2, 3, \dots, L$.

Observe that from (3.35) we can derive

$$(3.68) \quad 2\bar{a}\rho_1 \sum_{t=1}^N z(t)^2 \leq 2\bar{a} \sum_{t=1}^N |z(t)| \cdot |\nu_\omega(t+d-1)| + C_{11}r(N)^{1-\varepsilon} \\ + d_0 d\tilde{r}(N) + C_{12} \quad (\text{a.s.}),$$

where $0 < C_{11}, C_{12} < \infty$. Using the fact that

$$(3.69) \quad 2 \sum_{t=1}^N |z(t)| \cdot |\nu_\omega(t+d-1)| \\ \leq \rho'_0 \sum_{t=1}^N z(t)^2 + \frac{1}{\rho'_0} \sum_{t=1}^N \nu_\omega(t+d-1)^2, \quad 0 < \rho'_0 \ll \rho_1$$

and statement (3) of Lemma 3.1, (3.68) gives

$$(3.70) \quad r(N) \leq C_{14} \max_{1 \leq \tau \leq N} \|\phi(\tau)\|^2 + C_{15}N, \quad 0 < C_{14}, C_{15} < \infty \quad (\text{a.s.}).$$

Relations (3.67) and (3.70) imply, for $t \in [\sigma_{k_m}, \tau_{k_m} + 1)$ and $t \notin \mathcal{C}_L$, that

$$(3.71) \quad z(t)^2 \leq 2\varepsilon_2 \max_{1 \leq \tau \leq t} \|\phi(\tau)\|^2 + \varepsilon_2 \left\{ C_{14} \max_{1 \leq \tau \leq t} \|\phi(\tau)\|^2 + C_{15}t \right\}^{1-\varepsilon} \quad (\text{a.s.}),$$

where constants C_{14} and C_{15} are independent of ε_2 . Since, from relations (A.5) and (A.6) in the Appendix,

$$(3.72) \quad \max_{1 \leq \tau \leq t} \|\phi(\tau)\|^2 \leq C_{16} \max_{1 \leq \tau \leq t} z(\tau)^2 + C_{17}l_0(t), \quad 0 < C_{16}, C_{17} < \infty,$$

where $l_0(t)$ is defined by (3.11), from (3.71) we obtain that

$$(3.73) \quad z(t)^2 \leq \varepsilon_2 C_{19} \max_{1 \leq \tau \leq t} z(\tau)^2 + C_{20}t^{1-\varepsilon} + C_{21}l_0(t), \quad 0 < C_{19}, C_{20}, C_{21} < \infty \quad (\text{a.s.}),$$

wherefrom it follows, for $t \in [\sigma_{k_m}, \tau_{k_m} + 1)$ and $t \notin \mathcal{C}_L$, that

$$(3.74) \quad z(t)^2 \leq \varepsilon_3 \max_{1 \leq \tau \leq t} z(\tau)^2 + l_1(t) \quad (\text{a.s.}),$$

where ε_2 is chosen so that $\varepsilon_3 = \varepsilon_2 C_{19} < 1$, and

$$(3.75) \quad l_1(t) = C_{20}t^{1-\varepsilon} + C_{21}l_0(t).$$

Suppose that first L' points from the set \mathcal{C}_L are $t_1 = \sigma_{k_m}, \dots, t_{L'} = \sigma_{k_m} + L' - 1$, where $0 \leq L' \leq L$. Since conclusion (3.64) is true also for $n = m$, from statement (1) of Lemma 3.1, we obtain that

$$(3.76) \quad z(\sigma_{k_m} + L' - 1)^2 \leq \beta(\sigma_{k_m} + L' - 1),$$

where

$$(3.77) \quad \beta(\sigma_{k_m} + L' - 1) = C_\theta^{L'} \alpha(\sigma_{k_m} - 1) + \sum_{i=0}^{L'-1} C_\theta^i l_0(\sigma_{k_m} + L' - 1 - i),$$

where $\alpha(\sigma_{k_m} - 1)$ is defined by (3.65) for $n = m$. Thus, from (3.74) by (3.76), we get, for $t < t_{L'+1}, t_{L'+1} \in \mathcal{C}_L$, that

$$(3.78) \quad z(t)^2 \leq \varepsilon_3 \max_{1 \leq \tau \leq t} z(\tau)^2 + \beta(\sigma_{k_m} + L' - 1) + l_1(t) \quad (\text{a.s.}),$$

wherefrom it follows, for $t < t_{L'+1}$, that

$$(3.79) \quad \max_{1 \leq \tau \leq t} z(\tau)^2 \leq \frac{\beta(\sigma_{k_m} + L' - 1) + l_1(t)}{1 - \varepsilon_3} \quad (\text{a.s.}).$$

Using the “exponentially growing rule” (3.10) from the previous relation, we obtain that

$$(3.80) \quad z(t_{L'+1})^2 \leq C_\theta \frac{\beta(\sigma_{k_m} + L' - 1) + l_1(t_{L'+1})}{1 - \varepsilon_3} + l_0(t_{L'+1}) \quad (\text{a.s.}).$$

Therefore, by (3.74) and (3.79), we conclude that

$$(3.81) \quad \begin{aligned} z(t)^2 &\leq \varepsilon_3 \max_{1 \leq \tau \leq t} z(\tau)^2 + \beta(\sigma_{k_m} + L' - 1) + l_1(t) \\ &\quad + C_\theta \frac{\beta(\sigma_{k_m} + L' - 1) + l_1(t_{L'+1})}{1 - \varepsilon_3} \\ &\quad + l_0(t_{L'+1}) \quad (\text{a.s.}), \end{aligned}$$

for $t < t_{L'+2} \in \mathcal{C}_L$. The previous relation implies that

$$(3.82) \quad \begin{aligned} \max_{1 \leq \tau \leq t} z(\tau)^2 &\leq \frac{\beta(\sigma_{k_m} + L' - 1) + l_1(t) + l_0(t_{L'+1})}{1 - \varepsilon_3} \\ &\quad + C_\theta \frac{\beta(\sigma_{k_m} + L' - 1) + l_1(t_{L'+1})}{(1 - \varepsilon_3)^2} \quad (\text{a.s.}), \end{aligned}$$

for $t < t_{L'+2}$. Once again, using (3.10), from the above inequality, we obtain that

$$(3.83) \quad \begin{aligned} z(t_{L'+2})^2 &\leq C_\theta^2 \frac{\beta(\sigma_{k_m} + L' - 1) + l_1(t_{L'+1})}{(1 - \varepsilon_3)^2} \\ &\quad + C_\theta \frac{\beta(\sigma_{k_m} + L' - 1) + l_1(t_{L'+2})}{1 - \varepsilon_3} \\ &\quad + C_\theta \frac{l_0(t_{L'+1})}{1 - \varepsilon_3} + l_0(t_{L'+2}) \quad (\text{a.s.}). \end{aligned}$$

By induction, we can derive

$$(3.84) \quad z(t_L)^2 \leq C_0(t_L) \quad (\text{a.s.}),$$

where

$$(3.85) \quad \begin{aligned} C_0(t_L) &= \sum_{i=1}^{L-L'} C_\theta^{L-L'-i+1} \frac{\beta(\sigma_{k_m} + L' - 1) + l_1(t_{L'+i})}{(1 - \varepsilon_3)^{L-L'-i+1}} \\ &\quad + \sum_{i=1}^{L-L'} C_\theta^{L-L'-i} \frac{l_0(t_{L'+i})}{(1 - \varepsilon_3)^{L-L'-i}}. \end{aligned}$$

Finally, from (3.74) and (3.84), we get, for $t < \tau_{k_m+1}$, that

$$(3.86) \quad z(t)^2 \leq \varepsilon_3 \max_{1 \leq \tau \leq t} z(\tau)^2 + l_1(t) + C_0(t_L) \quad (\text{a.s.}),$$

wherefrom it follows that

$$(3.87) \quad \max_{1 \leq \tau \leq t} z(\tau)^2 \leq \frac{C_0(t_L) + l_1(t)}{1 - \varepsilon_3} \quad \text{for } t < \tau_{k_m+1} \quad (\text{a.s.}),$$

where $l_1(t)$ and $C_0(t_L)$ are defined by (3.75) and (3.85), respectively. Observe that the upper bound for $\max_{1 \leq \tau \leq t} z(\tau)^2$ in (3.87) is similar to the one established in (3.66). Now we are ready to establish mean-square boundedness of the residual $z(t)$.

Since $W_0(\sigma_k + d - 1) \leq 0$, and inequality (3.63) holds for all σ_k , we derive

$$(3.88) \quad \sum_{j=1}^t z(j)^2 \leq \frac{16}{\rho_1^2} \sum_{j=1}^{\sigma_k-1} \nu_\omega(j+d-1)^2 + \sum_{j=\sigma_k}^t z(j)^2.$$

Observe that from (3.62), for $t \in [\sigma_k, \tau_{k+1})$,

$$(3.89) \quad \sum_{j=\sigma_{k+1}}^t z(j)^2 \leq \frac{C_9}{\bar{a}\rho_1} \tilde{r}(t) \quad (\text{a.s.}).$$

Using the definition of $\tilde{r}(t)$ (3.6) and relation (3.70) from (3.89), we conclude, for $t \in [\sigma_k, \tau_{k+1})$, that

$$(3.90) \quad \sum_{j=\sigma_k}^t z(j)^2 \leq (C_{22} + 1) \max_{1 \leq \tau \leq t} z(\tau)^2 + C_{23}t^{1-\varepsilon} + C_{24}l_0(t) \quad (\text{a.s.}),$$

where $0 < C_{22}, C_{23}, C_{24} < \infty$. Combining (3.88) and (3.90), we get, for $t \in [\sigma_k, \tau_{k+1})$, that

$$(3.91) \quad \begin{aligned} \sum_{j=1}^t z(j)^2 &\leq \frac{16}{\rho_1^2} \sum_{j=1}^{\sigma_k-1} \nu_\omega(j+d-1)^2 \\ &\quad + \frac{16}{\rho_1^2} \left\{ C_{25} \max_{1 \leq \tau \leq t} z(\tau)^2 + C_{26}t^{1-\varepsilon} + C_{27}l_0(t) \right\} \quad (\text{a.s.}), \end{aligned}$$

where $0 < C_{25}, C_{26}, C_{27} < \infty$ and $l_0(t)$ is defined by (3.11).

Using the Markov inequality together with assumption (A_2) , we conclude that

$$(3.92) \quad \begin{aligned} \sum_{t=1}^{\infty} P\{\omega(t+1)^2 \geq t^c \mid \mathcal{F}_t\} \\ \leq \sum_{t=1}^{\infty} \frac{E\{\omega(t+1)^{2+\eta} \mid \mathcal{F}_t\}}{t^{(2+\eta)c/2}} < \infty \quad (\text{a.s.}), \end{aligned}$$

where $c \in (2/(2+\eta), 1)$. From the previous inequality by the conditional Borel–Cantely lemma (e.g., [19]) we derive, for all $c \in (2/(2+\eta), 1)$,

$$(3.93) \quad \omega(t+1)^2 = \mathcal{O}(t^c) \quad (\text{a.s.}),$$

wherefrom it follows that

$$(3.94) \quad \lim_{t \rightarrow \infty} \frac{l_0(t)}{t} = 0 \quad (\text{a.s.})$$

for $l_0(t)$ given by (3.11). Consequently, from (3.66), (3.87), and (3.94), it follows that

$$(3.95) \quad \begin{aligned} \limsup_{k \rightarrow \infty} \sup_k \sup_{t \in [\sigma_k, \tau_{k+1})} \frac{\max_{1 \leq \tau \leq t} z(\tau)^2}{t} \\ \leq C_{28} \limsup_{k \rightarrow \infty} \sup_k \sup_{t \in [\sigma_k, \tau_{k+1})} \frac{1}{t} \sum_{j=1}^{\sigma_k-1} \nu_\omega(j+d-1)^2, \end{aligned}$$

$$0 < C_{28} < \infty \quad (\text{a.s.}).$$

Finally, from (3.88), (3.89), (3.94), and (3.95), we conclude that

$$\begin{aligned}
 (3.96) \quad & \limsup_{k \rightarrow \infty} \sup_k \sup_{t \in [\sigma_k, \tau_{k+1})} \frac{1}{t} \sum_{j=1}^t z(j)^2 \\
 & \leq \frac{16}{\rho_1^2} (1 + C_2) \limsup_{k \rightarrow \infty} \sup_k \sup_{t \in [\sigma_k, \tau_{k+1})} \frac{1}{t} \sum_{j=1}^t \nu_\omega(j + d - 1)^2 \\
 & \leq (1 + C_2) \Sigma_0, \quad 0 < C_2 < \infty \quad (\text{a.s.}),
 \end{aligned}$$

where Σ_0 is defined by (3.39). Since in the time intervals $[\tau_k, \sigma_k)$, $W_0(t + d) \leq 0$, we conclude that relations (3.51)–(3.53) hold for $N \in [\tau_k, \sigma_k)$. From (3.53) it follows that

$$\limsup_{k \rightarrow \infty} \sup_k \sup_{t \in [\tau_k, \sigma_k)} \frac{1}{t} \sum_{j=1}^t z(j)^2 \leq \Sigma_0 \quad (\text{a.s.}),$$

where Σ_0 is given by (3.39). Thus statement (1) of the theorem is proved. Statements (2) and (3) of the theorem are direct consequences of relation (3.96), so the proof is complete. \square

The results presented in this section show that in the presence of unmodeled dynamics and external disturbances, the adaptive control algorithm possesses not only self-tuning, but also a self-stabilization property. The latter means the following: Whenever, as a consequence of incorrect parameter estimates, the adaptive system becomes unstable, the adaptive algorithm will stabilize itself by generating correct parameter estimates. During its operation, the adaptive controller passes through two phases characterized by the time intervals $[\tau_k, \sigma_k)$ and $[\sigma_k, \tau_{k+1})$, defined by (3.55). In the time intervals $[\tau_k, \sigma_k)$, $W_0(t + 1) \leq 0$, which implies the stability of the input and output signals for $t \in [\tau_k, \sigma_k)$. From (3.47) it is clear that in these time intervals no characterization of the function $V(t + 1)$ can be made. This function may diverge, thereby generating drifts of the parameter estimates. Consequently, controller parameters may escape from the set of stabilizing controllers, and the adaptive system will become unstable. Accordingly, the time intervals $[\tau_k, \sigma_k)$ correspond to the *drift* phase of the adaptive algorithm. As time progresses, the residual $z(t)$ becomes larger than $\nu_\omega(t)$, and the function $W_0(t + 1)$ given by (3.23) becomes positive. From relation (3.55), it is obvious that these periods of operation of the adaptive system correspond to the time intervals $[\sigma_k, \tau_{k+1})$, $k \geq 1$. Therefore drift of the parameter estimates in the time intervals $[\tau_k, \sigma_k)$ gives rise to the bursting phenomenon. The behavior of the Lyapunov function $V_1(t) = V(t) + [S(t) + W_0(t) + W_1(t) + W_2(t)]/\tilde{r}(t - d)$ is described by relation (3.48), and it is clear that $V_1(t + 1)$ decreases for $t \in [\sigma_k, \tau_{k+1})$. From (3.48) it also follows that in the time intervals $[\sigma_k, \tau_{k+1})$, fast adaptation takes place. Consequently, the parameter estimates reenter the set of stabilizing controllers. It is obvious that the time intervals $[\sigma_k, \tau_{k+1})$, $k \geq 1$, correspond to the *self-stabilization* phase of the adaptive system. To stabilize the system faster over the time intervals $[\sigma_k, \tau_{k+1})$, the algorithm gains μ are not required to be small. Specifically, a large μ allows for stronger influence of large tracking errors on the estimation of the controller parameters. On the other hand, to slow down the parameter drift in the time intervals $[\tau_k, \sigma_k)$, a small μ is required. The problem is that the bursting function $W_0(t + 1)$ is not measurable, and, consequently, the designer cannot change the coefficient μ during the operation of the adaptive system. It should be observed that we did not require the regressor $\phi(t)$ (or reference signal $y^*(t)$) to be persistently exciting.

Remark 1. It is not difficult to conclude that the boundedness of the parameter estimates is crucial to establishing the global stability of the adaptive system. The fact

that the parameter estimates are bounded makes it possible to establish an “exponentially growing rule,” which is stated in (3.10). This means that the signals in the adaptive loop cannot grow faster than exponentially. Also, relation (3.62) is obtained from (3.56) by exploiting the boundedness of the parameter estimates. As is shown in [15], in the case of the perfect system model, we do not need to bound parameter estimates by incorporating a projection in the estimation algorithm.

Remark 2. Observe that in our derivations we do not introduce any assumptions regarding the correlation of $\omega(t)$ and $y^*(t)$. They may be correlated.

Remark 3 (high-gain feedback). Let us consider the case where $d = 1$. Note that the error equation (2.7) can be written in the form

$$(3.97) \quad \begin{aligned} \tilde{C}(q^{-1})z(t) = & B(q^{-1})u(t) + [G(q^{-1}) + S(q^{-1})]y(t) - [C(q^{-1}) - 1]y^*(t) \\ & - y^*(t+1) - S(q^{-1})y(t) + \gamma(t) \\ & + [C(q^{-1}) - \tilde{C}(q^{-1})]\omega(t+1), \end{aligned}$$

i.e.,

$$(3.98) \quad \begin{aligned} \tilde{C}(q^{-1})z(t) = & \theta_0^T \phi(t) - y^*(t+1) + \gamma(t) - S(q^{-1})y(t) \\ & + [C(q^{-1}) - \tilde{C}(q^{-1})]\omega(t+1), \end{aligned}$$

where $S(q^{-1}) = s_0 + s_1 q^{-1} + \cdots + s_{n_s} q^{-n_s}$, $n_s = n_G$, and

$$(3.99) \quad \theta_0^T = \{\text{coeff}[G(q^{-1}) + S(q^{-1})]; \text{coeff } B(q^{-1}); \text{coeff}[\tilde{C}(q^{-1}) - 1]\},$$

while $\phi(t)$ is given by (3.3). Instead of (2.2), we assume for $\gamma(t)$ the following condition: Some polynomial $S(q^{-1})$ exists so that

$$(3.100) \quad |\gamma(t) - S(q^{-1})y(t)| \leq \gamma \sum_{j=1}^t \lambda_\gamma^{t-j} (|y(j)| + |\omega(j)| + k_\gamma),$$

$$\gamma > 0, \quad 0 < k_\gamma < \infty, \quad 0 < \lambda_\gamma < 1.$$

It is not difficult to see that all conclusions derived in § 3 are also valid in the case when $\gamma(t)$ satisfies (3.100), if parameter γ meets assumption (A₅). This means that, generally, for a given $\gamma(t)$, the proposed adaptive controller can result in high-gain feedback. In other words, to neutralize a large $\gamma(t)$, the algorithm will converge to a large s_i , $i = 0, 1, \dots, n_s$.

Remark 4 (nonlinear adaptive control). Using previous informations about the system, let us choose the functions that are nonlinear with respect to $y(t)$, given by

$$\phi_1(t) = \phi_1[y(t), y(t-1), \dots, y(1)], \dots, \phi_l(t) = \phi_l[y(t), y(t-1), \dots, y(1)],$$

so that

$$(3.101) \quad |\phi_j[y(t), y(t-1), \dots, y(1)]| \leq K_{\gamma j} \sum_{i=1}^t \lambda_{\gamma j}^{t-i} |y(i)| + k_{\phi j},$$

$$0 < K_{\gamma j}, k_{\phi j} < \infty, \quad 0 < \lambda_{\gamma j} < 1.$$

Note that for $d = 1$, (2.7) can be written in the form

$$(3.102) \quad \begin{aligned} \tilde{C}(q^{-1})z(t) = & \theta_0^T \phi(t) - y^*(t+1) + \gamma(t) - \sum_{j=1}^l \beta_j \phi_j(t) \\ & + [C(q^{-1}) - \tilde{C}(q^{-1})]\omega(t+1), \end{aligned}$$

where

$$(3.103) \quad \theta_0^T = \{\text{coeff } G(q^{-1}); \text{coeff } B(q^{-1}); \beta_1, \dots, \beta_l; \text{coeff } [C(q^{-1}) - 1]\}$$

and

$$(3.104) \quad \begin{aligned} \phi(t)^T = & [y(t), \dots, y(t-n_1); u(t), \dots, u(t-n_2); \\ & \phi_1(t), \dots, \phi_l(t); -y^*(t), \dots, -y^*(t-n_3+1)]. \end{aligned}$$

Let us suppose that parameters β_1, \dots, β_l exists, so that

$$(3.105) \quad \left| \gamma(t) - \sum_{j=1}^l \phi_j(t) \beta_j \right| \leq \gamma \sum_{j=1}^l \lambda_\gamma^{t-j} (|y(j)| + |\omega(j)| + k_\gamma),$$

$$\gamma > 0, \quad 0 < \lambda_\gamma < 1, \quad 0 < k_\gamma < \infty.$$

If γ satisfies condition (A_5) , repeating all derivations as in § 3, we can see that Theorem 3.1 is also true for $\gamma(t)$ given by (3.105).

Remark 5 (continuity property of the result formulated in Theorem 3.1). Suppose that $C(q^{-1}) - \bar{a}/2$ is a strictly positive real function. Consequently, $\tilde{C}(q^{-1}) = C(q^{-1})$ and the mean-square tracking error bound Σ_0 given by (3.39) is continuous with respect to γ . When $\gamma \rightarrow 0$, then $\Sigma_0 \rightarrow 0$ and the mean-square tracking error tends to the global minimum.

Remark 6 (positive realness of the system noise dynamics is not a necessary condition for the global convergence). Our result shows that the proposed algorithm operates even in the case when $C(z^{-1})$ is not a positive real function. Namely, it is well known that the estimation algorithm will operate if it correctly estimates the gradient of the functional criterion (2.4). If $C(z^{-1})$ is a strictly positive real function, then $\phi(t)(y(t+d) - y^*(t+d))$ is a good estimation of the gradient. If $C(z^{-1})$ is not strictly positive real, the proposed algorithm separates the strictly positive real part $\tilde{C}(z^{-1})$ from $C(z^{-1})$, and the residual part $\tilde{C}(z^{-1}) - C(z^{-1})$ generates the term $(C(q^{-1}) - \tilde{C}(q^{-1}))\omega(t)$. This term is treated by the algorithm as an external mean-square bounded disturbance, producing a larger upper bound for the mean-square tracking error (see (3.39)).

4. Output error method. Global stability properties of output error estimation schemes can be guaranteed under a strict positive realness (SPR) condition, but for many applications, especially in control, this condition is too restrictive. In recent years, several results have relaxed this condition.

In discrete time, the three results of interest are a global stability condition given in [20], a local stability/instability boundary presented in [14], and a concept of “composite regressor algorithm” proposed in [21]. In the context of the parameter identification problem in [20], it is shown that discrete-time output error algorithm in the case of a non-SPR condition is globally stable, provided that the plant input signal is persistently exciting and that the adaption gain times the input energy is sufficiently high where the input energy is the magnitude squared of the input signal. In the same context, “composite regressor algorithm,” an excellent idea is proposed in [21]. It enables the relaxation of the SPR condition, by a corresponding choice of a suitable design parameter. On the other hand, even in the ideal SPR case, in the case of output error method, the adaptive controller does not provide direct feedback from the system output to the input. There is only indirect influence of the system output through the controllers parameters estimates. Using the concept of composite regressor algorithm makes possible direct output feedback, thereby enabling us to avoid the SPR condition.

In this section, using the above-mentioned idea, we consider stochastic adaptive control problem and prove global stability (by methodology developed in § 3) without imposing the SPR condition on the system dynamics.

Let us consider the system model

$$(4.1) \quad (1 - A(q^{-1}))y(t) = B(q^{-1})u(t-1),$$

where polynomial $A(q^{-1})$ is defined by

$$(4.2) \quad A(q^{-1}) = a_1 q^{-1} + \cdots + a_{n_A} q^{-n_A},$$

while $B(q^{-1})$ is the same as in (2.2). We suppose that system output $y(t)$ is corrupted by colored noise $\xi(t)$ given by

$$(4.3) \quad P(q^{-1})\xi(t) = Q(q^{-1})\omega(t);$$

i.e., the measurable variable is

$$(4.4) \quad y_M(t) = y(t) + \xi(t).$$

In (4.3), $\omega(t)$ is the martingale difference sequence defined by assumption (A₂), while polynomials $P(q^{-1})$ and $Q(q^{-1})$ are defined as follows:

$$P(q^{-1}) = 1 + p_1 q^{-1} + \cdots + p_{n_P} q^{-n_P}$$

and

$$(4.5) \quad Q(q^{-1}) = 1 + q_1 q^{-1} + \cdots + q_{n_Q} q^{-n_Q}.$$

We introduce the following assumption concerning noise dynamic:

$$(A_6) \quad P(z^{-1}) \text{ has zeros strictly outside the unit disc.}$$

Similarly, as in § 2, in the case of the unknown systems parameters, we seek to design an adaptive controller so that the functional criterion (2.3) is minimized.

From (4.1) and (4.4), we can get that

$$(4.6) \quad \begin{aligned} & \{1 - (1 - \alpha)A(q^{-1})\}(y_M(t+1) - y^*(t+1) - \xi(t+1)) \\ &= B(q^{-1})u(t) + \alpha A(q^{-1})y_M(t) \\ & \quad + (1 - \alpha)A(q^{-1})y^*(t) - y^*(t+1) - \alpha A(q^{-1})\xi(t), \end{aligned}$$

where $0 \leq \alpha < 1$ is the design parameter. Using the concept of the “composite regressor” [21], we define composite output “prediction”

$$(4.7) \quad \bar{y}(t) = \alpha y_M(t) + (1 - \alpha)y^*(t).$$

Equation (4.6) can now be rewritten in the following form:

$$(4.8) \quad \begin{aligned} & \{1 - (1 - \alpha)A(q^{-1})\}(y_M(t+1) - y^*(t+1) - \xi(t+1)) \\ &= \theta_0^T \phi(t) - y^*(t+1) - \alpha A(q^{-1})\xi(t), \end{aligned}$$

where

$$(4.9) \quad \theta_0^T = [a_1, \cdots, a_{n_A}; b_0, \cdots, b_{n_B}]$$

and

$$(4.10) \quad \phi(t)^T = [\bar{y}(t), \cdots, \bar{y}(t - n_A + 1); u(t), \cdots, u(t - n_B)].$$

For the purpose of our analysis, we rearrange (4.8) in the form

$$(4.11) \quad \{1 - (1 - \alpha)A(q^{-1})\}z(t) = \theta_0^T \phi(t) - y^*(t+1) + \gamma(t),$$

where

$$(4.12) \quad z(t) = y_M(t+1) - y^*(t+1) - \omega(t+1)$$

and

$$(4.13) \quad \gamma(t) = \{1 - A(q^{-1})\} \left\{ \frac{Q(q^{-1})}{P(q^{-1})} - 1 \right\} \omega(t+1) - \alpha A(q^{-1}) \omega(t+1).$$

In the subsequent analysis, we assume that assumption (A₄) introduced in § 3 is also valid for θ_0 , defined by (4.9). For the estimation of the unknown parameters vector θ_0 , we use the following algorithm:

$$(4.14) \quad \hat{\theta}(t+1) = F \left\{ \hat{\theta}(t) + \frac{\bar{a}\phi(t)}{\tilde{r}(t)} [y_M(t+1) - y^*(t+1)] \right\}, \quad 0 < \bar{a} < 1,$$

where $F\{\cdot\}$ is the projection operator as in (3.5), and $\tilde{r}(t)$ is defined by (3.6) for $\phi(t)$ given by (4.10). Adaptive control law is

$$(4.15) \quad \hat{\theta}(t)^T \phi(t) = y^*(t+1),$$

where $\hat{\theta}(t)$ is the estimate of θ_0 , obtained by the recursive scheme (4.14). From (4.7) and (4.15), it is obvious that larger α implies larger output feedback and more chance for $1 - (1 - \alpha)A(z^{-1})$ to be an SPR function.

Since α is the design parameter with its choice function, $1 - (1 - \alpha)A(z^{-1})$ can be made strictly positive real. If α is closer to 1, the chances are higher that $1 - (1 - \alpha)A(z^{-1})$ is SPR. It is well known that $1 - (1 - \alpha)A(z^{-1}) - \bar{a}/2$, where $0 < \bar{a} < 1$, defined in (4.14), is SPR if

$$(4.16) \quad (1 - \alpha) \sum_{j=1}^{n_A} |a_j| < 1 - \frac{\bar{a}}{2},$$

where $a_j, j = 1, \dots, n_A$ are the coefficients of the polynomial $A(q^{-1})$.

Observe that assumption (A₄) implies that the upper bound $d_\theta < \infty$ of θ_0 is known ($\|\theta_0\| \leq d_\theta < \infty$). If we choose α so that inequality $1 - (1 - \bar{a}/2)/d_\theta < \alpha < 1$ holds, relation (4.16) will be satisfied, and, consequently, function $1 - (1 - \alpha)A(z^{-1}) - \bar{a}/2$ will be SPR. Thus we can assume that the following holds:

(A₇) Parameter α is chosen so that $1 - (1 - \bar{a}/2)/d_\theta < \alpha \leq 1$.

Since (A₇) implies strict positive realness property of the function $1 - (1 - \alpha)A(z^{-1}) - \bar{a}/2$, we can formulate the following global stability result.

THEOREM 4.1. *Let assumptions (A₁)-(A₄), (A₆), and (A₇) hold. Then*

$$(4.17) \quad (1) \quad \lim_{N \rightarrow \infty} \sup_N \frac{1}{N} \sum_{t=1}^N z(t)^2 \leq \Sigma'_0 \quad (\text{a.s.}),$$

where

$$(4.18) \quad \Sigma'_0 = \frac{1 + C'_2}{2(\bar{a}\rho_m - \rho_0)} C_\alpha \sigma_\omega^2,$$

where $0 < C'_2 < \infty$, $0 < \rho_0 \ll \bar{a}\rho_m$, and $\rho_m > 0$ is the largest number, so that $1 - (1 - \alpha)A(z^{-1}) - \bar{a}/2 - \rho_m$ remains a positive real function, while

$$(4.19) \quad C_\alpha = \max_{|z|=1} \left| (1 - A(z)) \left(\frac{Q(z)}{P(z)} - 1 \right) - \alpha A(z) \right|^2,$$

$$(4.20) \quad (2) \quad \lim_{N \rightarrow \infty} \sup_N \frac{1}{N} \sum_{t=1}^N (y(t) - y^*(t))^2 \leq \Sigma'_0 + 2(\Sigma'_0 C_\beta)^{1/2} + C_\beta \quad (\text{a.s.}),$$

where

$$(4.21) \quad C_\beta = \max_{|z|=1} \left| 1 - \frac{Q(z)}{P(z)} \right|^2,$$

$$(4.22) \quad (3) \quad \lim_{N \rightarrow \infty} \sup_N \frac{1}{N} \sum_{t=1}^N \|\phi(t)\|^2 \leq C'_1 < \infty \quad (\text{a.s.}).$$

Proof. Starting from (4.14), we can get that

$$(4.23) \quad \begin{aligned} V(t+1) \leq & V(t) + \frac{2\bar{a}}{\tilde{r}(t)} \tilde{\theta}(t)^T \phi(t) z(t) + \frac{2\bar{a}}{\tilde{r}(t)} \tilde{\theta}(t)^T \phi(t) \omega(t+1) \\ & + \frac{2\bar{a}^2 \|\phi(t)\|^2}{\tilde{r}(t)^2} \{z(t)^2 + \omega(t+1)^2\}, \end{aligned}$$

where $V(t) = \|\tilde{\theta}(t)\|^2$, $\tilde{\theta}(t) = \hat{\theta}(t) - \theta_0$ and $z(t)$ is defined by (4.12).

Since, from (4.11) and (4.15),

$$(4.24) \quad \{1 - (1 - \alpha)A(q^{-1})\}z(t) = -\tilde{\theta}(t)^T \phi(t) + \gamma(t),$$

after simple majorizations from (4.23), we obtain that

$$(4.25) \quad \begin{aligned} V(t+1) \leq & V(t) - \frac{2\bar{a}}{\tilde{r}(t)} \left\{ \left[1 - (1 - \alpha)A(q^{-1}) - \frac{\bar{a}}{2} \right] z(t) \right\} z(t) \\ & + \frac{2\bar{a}|z(t)| |\gamma(t)|}{\tilde{r}(t)} + \frac{2\bar{a}}{\tilde{r}(t)} \tilde{\theta}(t)^T \phi(t) \omega(t+1) \\ & + \frac{2\bar{a}^2 \|\phi(t)\|^2}{\tilde{r}(t)^2} \omega(t+1)^2. \end{aligned}$$

Let us define the sequence

$$(4.26) \quad S_1(t+1) = \sum_{j=1}^t z(j) \left\{ 1 - (1 - \alpha)A(q^{-1}) - \frac{\bar{a}}{2} - \rho_m \right\} z(j) + K'_3,$$

$$0 < K'_3 < \infty.$$

If we select $\rho_m > 0$ as the largest number so that $1 - (1 - \alpha)A(z^{-1}) - \bar{a}/2 - \rho_m$ remains positive real, then $S_1(t+1) \geq 0$ for all $t \geq 0$. The existence of such ρ_m is ensured by assumption (A₇). It provides strict positive realness of the function $1 - (1 - \alpha)A(z^{-1}) - \bar{a}/2$. Using this fact, we can write relation (4.25) in the form

$$(4.27) \quad \begin{aligned} V(t+1) + \frac{S_1(t+1)}{\tilde{r}(t)} \leq & V(t) + \frac{S_1(t)}{\tilde{r}(t)} - 2\bar{a}\rho_m \frac{z(t)^2}{\tilde{r}(t)} \\ & + \frac{2\bar{a}|z(t)| |\gamma(t)|}{\tilde{r}(t)} + \frac{2\bar{a}\tilde{\theta}(t)^T \phi(t) \omega(t+1)}{\tilde{r}(t)} \\ & + \frac{2\bar{a}^2 \|\phi(t)\|^2}{\tilde{r}(t)^2} \omega(t+1)^2. \end{aligned}$$

The rest of the proof is similar to the proof of Theorem 3.1 and will be omitted. \square

From (4.24), we see that $\gamma(t)$ acts as an external mean-square bounded disturbance. The presence of this term makes possible bursting phenomena to appear. Self-stabilization mechanism described in § 3 takes place in this case as well, thus ensuring a globally stable closed-loop system.

Remark 7. Note that the number $\rho_m, \rho_m < 1 - \bar{a}/2$ can be predefined by choosing α so that

$$(4.28) \quad 1 - \frac{1 - (\bar{a}/2) - \rho_m}{d_\theta} \leq \alpha \leq 1.$$

Then

$$(4.29) \quad (1 - \alpha) \sum_{j=1}^{n_A} |a_j| \leq (1 - \alpha) d_\theta \leq 1 - \frac{\bar{a}}{2} - \rho_m,$$

and, consequently, the function $1 - (1 - \alpha)A(z^{-1}) - \bar{a}/2 - \rho_m$ is positive real that satisfies the condition of $S_1(t)$ defined by (4.26) to be nonnegative.

Remark 8. Based on (4.18), it follows that when α is closer to 1, then a higher bound for mean-square tracking error is obtained.

Remark 9. The only nonideal characteristic of system (4.1) is that the output error transfer function is not SPR. When in the system model (4.1) there is unmodeled dynamics like the one defined by relation (2.3), global stability proof can be easily handled by using the methodology developed in § 3.

5. Conclusions. A new algorithm for robust stochastic adaptive control in the presence of the unmodeled dynamics has been proposed. The self-stabilization mechanism that guarantees global stability of the closed-loop system has been evaluated analytically. The algorithm results in a minimum-variance self-tuning control when the unmodeled dynamics disappear. The methodology underlying the algorithm is presently applied to adaptive filtering and prediction when modeling errors are considered. By identifying interconnections among subsystems as unmodeled dynamics, this methodology is being used for decentralized control, identification, and computation in complex systems.

Appendix.

Proof of Lemma 3.1. Since $\tilde{C}(q^{-1})$ is a stable operator, from (3.2) we can derive

$$(A.1) \quad z(t)^2 \leq k_0 \max_{1 \leq \tau \leq t} \|\phi(\tau)\|^2 + m'_1 + \max_{1 \leq \tau \leq t} \nu(\tau + d - 1)^2, \quad 0 < k_0, m'_1 < \infty.$$

Observe that

$$(A.2) \quad \gamma(t + d - 1)^2 \leq k_1 \max_{1 \leq \tau \leq t} z(\tau - 1)^2 + k_2 \max_{1 \leq \tau \leq t} \omega(\tau + d - 1)^2 + k_3, \\ 0 < k_1, k_2, k_3 < \infty$$

and

$$(A.3) \quad \max_{1 \leq \tau \leq t} \|\phi(\tau)\|^2 \leq (n_1 + 1) \max_{1 \leq \tau \leq t} y(\tau)^2 + n_2 \max_{1 \leq \tau \leq t} u(\tau - 1)^2 \\ + \max_{1 \leq \tau \leq t} u(\tau)^2 + n_3 m_1^2.$$

From (3.8), by assumption (A₄), we have that

$$(A.4) \quad \max_{1 \leq \tau \leq t} u(\tau)^2 \leq k_4 \max_{1 \leq \tau \leq t} u(\tau - 1)^2 + k_5 \max_{1 \leq \tau \leq t} y(\tau)^2 + k_6, \quad 0 < k_4, k_5, k_6 < \infty,$$

wherefrom by (A.3) we conclude that

$$(A.5) \quad \max_{1 \leq \tau \leq t} \|\phi(\tau)\|^2 \leq k_7 \max_{1 \leq \tau \leq t} y(\tau)^2 + k_8 \max_{1 \leq \tau \leq t} u(\tau-1)^2 + k_9, \\ 0 < k_7, k_8, k_9 < \infty.$$

From (2.1), by using assumption (A₁), we derive

$$(A.6) \quad \max_{1 \leq \tau \leq t} u(\tau-1)^2 \leq k_{10} \max_{1 \leq \tau \leq t} y(\tau+d-1)^2 + k_{11} \max_{1 \leq \tau \leq t} \omega(\tau+d-1)^2 \\ + k'_{11} \max_{1 \leq \tau \leq t} \gamma(\tau+d-2)^2, \\ 0 < k_{10}, k_{11}, k'_{11} < \infty.$$

Combining (A.1), (A.2), (A.5), and (A.6), we obtain that

$$(A.7) \quad z(t)^2 \leq C_\theta \max_{1 \leq \tau \leq t} z(\tau-1)^2 + k'_\theta \max_{1 \leq \tau \leq t} \omega(\tau+d-1)^2 + k''_\theta, \\ 0 < C_\theta, k'_\theta, k''_\theta < \infty,$$

by which statement (1) of the lemma is proved.

Statements 2(i) and 2(ii) follow directly from the definitions of the sequences $z(t)$, $\gamma(t)$, and $\nu(t)$. Statement 2(iii) is a consequence of the following simple inequalities:

$$(A.8) \quad \sum_{t=1}^N \nu_z(t+d-1)^2 \leq (n_F+1) \sum_{i=0}^{n_F} f_i^2 \sum_{t=1}^N \gamma_z(t+d-1)^2,$$

$$(A.9) \quad \sum_{t=1}^N \gamma_z(t+d)^2 \leq \frac{\gamma^2}{(1-\lambda_\gamma)^2} \sum_{t=1}^N z(t)^2 + \xi_1,$$

and

$$(A.10) \quad \sum_{t=1}^N |z(t)| \cdot |\nu_z(t+d-1)| \leq \left(\sum_{t=1}^N z(t)^2 \right)^{1/2} \left(\sum_{t=1}^N \nu_z(t+d-1)^2 \right)^{1/2},$$

where $\gamma_z(t)$ and $\nu_z(t)$ are given by (3.13) and (3.16), respectively. Let us prove statement (3) of our lemma.

Observe that

$$(A.11) \quad \sum_{t=1}^N \|\phi(t)\|^2 \leq (n_1+1) \sum_{t=1}^N y(t)^2 + (n_2+1) \sum_{t=1}^N u(t)^2 + n_3 m_1^2 N$$

and, from (2.1), by “stable invertibility” assumption (A₁),

$$(A.12) \quad \sum_{t=1}^N u(t)^2 \leq 3C_{BA}^2 \sum_{t=1}^N y(t+d)^2 + 3C_{BC}^2 \sum_{t=1}^N \omega(t+d)^2 \\ + 3C_B^2 \sum_{t=1}^N \gamma(t+d-1)^2,$$

where C_{BA} , C_{BC} , and C_B are defined by (3.9). Since, from (2.2),

$$(A.13) \quad \sum_{t=1}^N \gamma(t+d-1)^2 \leq \frac{3\gamma^2}{(1-\lambda_\gamma)^2} \sum_{t=1}^N y(t+d-1)^2 \\ + \frac{3\gamma^2}{(1-\lambda_\gamma)^2} \sum_{t=1}^N \omega(t+d)^2 + \frac{3\gamma^2 k_\gamma^2}{(1-\lambda_\gamma)^2} N,$$

by (A.12) we can get that

$$(A.14) \quad \sum_{t=1}^N u(t)^2 \leq 3 \left[C_{BA}^2 + \frac{3\gamma^2 C_B^2}{(1-\lambda_\gamma)^2} \right] \sum_{t=1}^N y(t+d)^2 + 3 \left[C_{BC}^2 + \frac{3\gamma^2 C_B^2}{(1-\lambda_\gamma)^2} \right] \cdot \sum_{t=1}^N \omega(t+d)^2 + \frac{9\gamma^2 C_B^2 k_\gamma^2}{(1-\lambda_\gamma)^2} N.$$

Since, by assumption (A₂),

$$(A.15) \quad \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{t=1}^N \omega(t)^2 = \sigma_\omega^2 \quad (\text{a.s.}),$$

using the definition of $z(t)$, by combining (A.14) and (A.15), from (A.11) we can easily obtain statement (3) of the lemma. Thus the lemma is proved. \square

LEMMA. Let assumption (A₂) hold. Then, for $1 \leq i < \infty$,

$$(A.16) \quad (1) \quad \sum_{t=1}^N \frac{\|\phi(t)\|^2}{\tilde{r}(t)(r(t)+t)^{2\varepsilon}} \omega(t+i)^2 \leq C_1 < \infty \quad (\text{a.s.}),$$

$$(A.17) \quad (2) \quad \sum_{t=1}^N \frac{\|\phi(t)\|^2}{\tilde{r}(t)} \omega(t+i)^2 \leq o\{(r(N)+N)^{2\varepsilon}\} \quad (\text{a.s.}),$$

where $\tilde{r}(t)$ and $r(t)$ are defined by (3.6) and (3.7), respectively, while $0 < \varepsilon < \frac{1}{2}$.

Proof. Let us define the following sequence:

$$(A.18) \quad D(N+i) = \sum_{t=1}^N \frac{\|\phi(t)\|^2}{\tilde{r}(t)(r(t)+t)^{2\varepsilon}} \omega(t+i)^2, \quad 1 \leq i < \infty.$$

Since

$$(A.19) \quad D\{D(N+i) | \mathcal{F}_{N+i-1}\} \leq D(N+i-1) + \frac{\|\phi(N)\|^2}{\tilde{r}(N)(r(N)+N)^{2\varepsilon}} \sigma_\omega^2$$

and

$$(A.20) \quad \sum_{t=1}^{\infty} \frac{\|\phi(t)\|^2}{r(t)^{1+\varepsilon}} \leq C_2 < \infty$$

by the martingale convergence theorem, we conclude that

$$(A.21) \quad \lim_{N \rightarrow \infty} D(N) = D < \infty \quad (\text{a.s.}).$$

Thus statement (1) of the lemma is proved. By Kronecker's lemma, from (A.21) we obtain that

$$(A.22) \quad \lim_{N \rightarrow \infty} \frac{1}{(r(N)+N)^{2\varepsilon}} \sum_{t=1}^N \frac{\|\phi(t)\|^2}{\tilde{r}(t)} \omega(t+i)^2 = 0 \quad (\text{a.s.}),$$

i.e.,

$$(A.23) \quad \sum_{t=1}^N \frac{\|\phi(t)\|^2}{\tilde{r}(t)} \omega(t+i)^2 \leq o((r(N)+N)^{2\varepsilon}) \quad (\text{a.s.}),$$

and the proof of the lemma is complete. \square

REFERENCES

- [1] C. E. ROHRS, L. VALAVANI, M. ATHANS, AND C. STEIN, *Analytical verification of undesirable properties of direct model reference adaptive control algorithm*, in Proc. 20th IEEE Conference on Decision and Control, San Diego, CA, 1981, pp. 1272–1284.
- [2] ———, *Robustness of adaptive control algorithms in the presence of unmodeled dynamics*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 881–889.
- [3] B. D. O. ANDERSON, R. R. BITMEAD, C. R. JOHNSON, JR., P. V. KOKOTOVIC, R. L. KOSUT, I. M. Y. MAREELS, L. PRALY, AND B. D. RIEDLE, *Stability of Adaptive Systems: Passivity and Averaging Analysis*, MIT Press, Cambridge, MA, 1986.
- [4] R. ORTEGA AND T. YU, *Theoretical results on robustness of direct adaptive controllers: A survey*, in Proc. 10th IFAC World Congress, Munich, Germany, 1987, pp. 26–31.
- [5] B. EGARDT, *Stability of Adaptive Controllers*, Springer-Verlag, Berlin, New York, 1979.
- [6] K. L. ASTROM, *Analysis of Rohrs counterexamples to adaptive control*, in Proc. American Control Conference, San Diego, CA, 1984, pp. 87–93.
- [7] B. D. RIEDLEE AND P. V. KOKOTOVIC, *Bifurcating equilibria of adaptive control systems*, in Proc. American Control Conference, San Diego, CA, 1984, pp. 238–241.
- [8] B. D. O. ANDERSON, *Adaptive systems, lack of persistency of excitation and bursting phenomena*, Automatica, 21 (1985), pp. 247–258.
- [9] I. M. Y. MAREELS AND R. R. BITMEAD, *Non-linear dynamics in adaptive control: Chaotic and periodic stabilization*, Automatica, 22 (1986), pp. 641–655.
- [10] B. E. YDSTIE, *Bifurcations and Complex Dynamics in Adaptive Control Systems*, in Proc. 25th IEEE Conference on Decision and Control, Athens, Greece, 1986, pp. 786–790.
- [11] M. JAIDANE-SAIDANE AND O. MACHI, *Quasiperiodic self-stabilization of adaptive ARMA predictors*, Internat. J. Adaptive Control Signal Process., 2 (1988), pp. 1–31.
- [12] L. PRALY AND J. B. POMET, *Periodic solutions in adaptive systems: The regular case*, in Proc. 10th IFAC World Congress on Automatic Control, Munich, Germany, 1987, pp. 166–171.
- [13] I. M. Y. MAREELS AND R. R. BITMEAD, *Bifurcation effects in robust adaptive control*, IEEE Trans. Circuits and Systems, 35 (1988), pp. 835–841.
- [14] D. A. SCHOENWALD AND P. V. KOKOTOVIC, *Boundedness conjecture for an output error adaptive algorithm*, Internat. J. Adaptive Control Signal Process., 4 (1990), pp. 27–47.
- [15] G. C. GOODWIN AND K. S. SIN, *Adaptive Filtering, Prediction and Control*, Prentice-Hall, Englewood Cliffs, NJ, 1984.
- [16] P. R. KUMAR AND P. VARAIYA, *Stochastic Systems: Estimation, Identification and Adaptive Control*, Prentice-Hall, Englewood Cliffs, NJ, 1986.
- [17] H. F. CHEN AND L. GUO, *A robust stochastic adaptive controller*, IEEE Trans. Automat. Control, 33 (1988), pp. 1035–1043.
- [18] L. PRALY, S. F. LIN, AND P. R. KUMAR, *A robust adaptive minimum variance controller*, SIAM J. Control Optim., 27 (1989), pp. 235–266.
- [19] W. F. STOUT, *Almost Sure Convergence*, Academic Press, New York, 1974.
- [20] M. TOMIZUKA, *Parallel MRAS without compensation block*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 505–506.
- [21] J. B. KENNEY AND C. E. ROHRS, *The composite regressor algorithm*, in IEEE ICASSP Proc., New York, 1988, pp. 1561–1563.

GLOBAL TIME-VARYING LINEARIZATION UP TO OUTPUT INJECTION*

H. HAMMOURI† AND J. P. GAUTHIER‡

Abstract. This paper, following the purpose of synthesis of observers for nonlinear systems, investigates the question of equivalence of a nonlinear system with a bilinear system, or, more generally, a linear time-dependent system, plus an output injection. In a previous work by the same authors, such questions have already been dealt with from the local point of view. The goal herein is to examine the global situation. Using basic facts from algebraic topology, it is shown that in the single output case, whenever the possibility to bilinearize up to output injection holds locally everywhere, it also holds globally.

Key words. output injection, bilinearization of nonlinear systems, fibre bundles

AMS(MOS) subject classifications. 93C10, 93B07

1. Introduction. We consider the following general nonlinear systems (Σ):

$$(\Sigma) \quad \dot{x} = f(x(t), u(t)) = f_u(x), \quad y(t) = h(x(t)),$$

where $x(t) \in M$ (a C^r n -dimensional connected manifold), $u(t) \in R^m$, $y(t) \in R$, and $\{f_u\}$ is a family of C^r vector fields on M . We deal only with the cases where $r = \infty$ or $r = \omega$.

We also consider linear time-varying systems up to output injection (\mathcal{L}), i.e., systems of the form

$$(\mathcal{L}) \quad \dot{z}(t) = A(u(t))z(t) + \varphi(u(t), y(t)), \quad y(t) = C \cdot z(t) = z_1(t),$$

where $A(u(t))$ is a matrix depending on the control u , $z(t) \in R^n$, $y(t) \in R$, C is a constant linear form, and $\varphi(u, y)$ a vector field depending on u and y only.

In the following, we assume that the considered (Σ) and (\mathcal{L}) satisfy the following observability assumptions:

(1) (Σ) and (\mathcal{L}) are observable: Given any two distinct initial conditions, there is a control function $u(\cdot)$ such that the associated outputs are different;

(2) (Σ) and (\mathcal{L}) satisfy the observability rank condition: Let σ be the observation space, i.e., the smallest real vector space containing the output map (h , respectively, C) and closed under Lie differentiation with respect to the dynamics of the system. Then the vector space $\{d\tau(x), \tau \in \sigma\}$ is n -dimensional for any $x \in M$ (respectively, $x \in R^n$).

Under these assumptions, we state the following definition.

DEFINITION 1. (i) (Σ) is “locally linearizable up to output injection” at $x_0 \in M$ if and only if there exists a diffeomorphism defined on some neighbourhood of x_0 onto an open subset of R_n that transforms (Σ) into a system of the form (\mathcal{L}). We write that (Σ) is L.L.O.I. at x_0 .

(ii) (Σ) is “everywhere locally linearizable up to output injection” if (Σ) is L.L.O.I. at each $x \in M$. We write E.L.L.O.I.

(iii) (Σ) is “everywhere locally linearizable up to output injection, with a global output injection” if (Σ) is E.L.L.O.I., and the output injection is global on M . We write E.L.L.O.I.G.

* Received by the editors December 22, 1989; accepted for publication (in revised form) June 26, 1991.

† Laboratoire d'Automatique et de Génie de Procédés (LAGEP), Unité de Recherche Associée/Centre National de la Recherche Scientifique D1328, Université Claude Bernard, Lyon 1, Bâtiment 721, 43 bd du 11 Novembre 1918, 69622 Villeurbanne Cedex, France.

‡ Laboratoire de Mathématique et Informatique, Institut National des Sciences Appliquées de Rouen, B.P. 8, F76131 Mont Saint-Aignan Cedex, France.

(iv) (Σ) is "globally linearizable up to output injection" if there exists a global diffeomorphism from M to an open subset of R^n that sends (Σ) to a system of the form (\mathcal{L}) . We write G.L.O.I.

Our purpose in this paper is to investigate whether E.L.L.O.I. is equivalent to G.L.O.I., especially in the single output case. The practical interest of this question is related to the problem of synthesis of observers. Following the suggestions of a referee, we explain this point in detail.

Several authors [LM], [K], [KI], [KR], [BRG] have investigated the case when a system can be transformed into a linear system, plus an output injection. Consider the following such system:

$$(\Sigma_1) \quad \dot{x} = Ax + bu + \varphi(y, u), \quad y = Cx.$$

(Σ_1) is observable independently of the value of the input u if and only if the linear pair (C, A) is observable. In that case, choosing K such that $(A - KC)$ has spectrum with negative real part, we get that

$$(\mathcal{O}_1) \quad \dot{\hat{x}} = A\hat{x} + bu + \varphi(y, u) - K(C\hat{x} - y)$$

is an observer system for (Σ_1) with exponential decay for the error estimate.

This property that a nonlinear system is observable independently of the input, true for observable linear systems, is highly nongeneric and hides the major difficulty in the observer synthesis problem: generally, the observability property is input-dependent. Also, generally, few inputs make the system unobservable. However, this is sufficient to render the observation problem a difficult problem: for inputs close to these bad inputs, the observation becomes more difficult; therefore we cannot expect to get observer systems in a strong sense such as in the linear case. Thus linear systems plus output injection are not representative of the main difficulty of the synthesis of observers.

On the contrary, bilinear systems, or, more generally, state affine systems, have generically bad inputs that make them unobservable. Despite this, they have nice observers that work for good inputs. For example, state affine systems, the inputs being known, are just linear time-dependent systems; hence the standard optimal Kalman's observer works as follows:

$$(\Sigma_2) \quad \dot{x} = A(u)x + bu, \quad x \in \mathbb{R}^n, \quad y = Cx,$$

$$(i) \quad \dot{\hat{x}} = A(u)\hat{x} + bu - S^{-1}C'(C\hat{x} - y),$$

$$(\mathcal{O}_2)$$

$$(ii) \quad \dot{S} = -A(u)'S - SA(u) + C'C - SQS,$$

where the constant Q and $S(t)$ remain in the cone of positive definite matrices.

It is known that as soon as $u(t)$ makes the linear time-dependent system (Σ_2) observable in some strong sense (i.e., $u(t)$ is a regularly persistent input in the sense defined in [GCKS]), (\mathcal{O}_2) is an observer system for (Σ_2) . See also [BCC] for details.

If we add to (Σ_2) an output injection

$$(\Sigma'_2) \quad \dot{x} = A(u)x + bu + \varphi(y, u), \quad y = Cx,$$

it is easily seen that $u(t)$ is a good input for (Σ'_2) (making (Σ'_2) observable) if and only if $u(t)$ is a good input for (Σ_2) , and that the modified observer

$$(\mathcal{O}'_2) \quad \dot{\hat{x}} = A(u)\hat{x} + bu - S^{-1}C'(C\hat{x} - y) + \varphi(y, u),$$

$$\dot{S} = -A(u)'S - SA(u) + C'C - SQS$$

works as soon as the input $u(t)$ is good (in the same strong sense as above). Other Kalman type observers are known, with an arbitrarily specified exponential decay of the error for good inputs; again, see [BCC].

Hence we claim that the problem of transforming a system into a state affine system plus an output-injection is a very important problem for the purpose of the synthesis of observers and that it is more representative of the main difficulties, since it allows the consideration of nonlinear systems having bad inputs.

Bilinearization (with or without additive output injection) has been treated by several authors; see the basic paper [FK] and also [GCKS], [HG1], [HG2].

The global questions, to our knowledge, have not yet been treated. However, a similar approach for feedback equivalence has been developed; see, for example, [D], [B]. These authors, of course, point out some relationship between these questions and standard techniques related to foliations (holonomy) and algebraic topology in the large.

Let us now recall the local result and state our main conclusions. Given a vector field X , we denote as usual by L_X , i_X , and d the Lie derivative with respect to X , interior derivative with respect to X , and exterior differentiation of p -forms. A nonlinear system (Σ) being given, and X being some vector field, we denote by Ω_1^X the real vector space generated by $\{dL_{f_u}(h) \wedge dh, u \in R^m\}$ (\wedge : standard exterior product). By induction, we define Ω_{k+1}^X as the real vector space generated by

$$\{L_{f_u}(i_X \omega) \wedge dh; \omega \in \Omega_k^X, u \in R^m\}.$$

We set $\Omega^X = \sum_{k=1}^{\infty} \Omega_k^X$. Note that, despite the superscript X , Ω_1^X does not depend on X .

We state the following theorem.

THEOREM 2 (see [HG1, p. 140]). *(Σ) is L.L.O.I. at $x_0 \in M$ if and only if*

- (1) $dh(x_0) \neq 0$;
- (2) *There exists a vector field X , defined on some neighbourhood of x_0 such that*
 - (a) $L_X h = 1$,
 - (b) $\dim_R(\Omega^X) = n - 1$,
 - (c) *for all $\omega \in \Omega^X$, $d(i_X \omega) = 0$,*
 - (d) $\Lambda^{n-1}(i_X \Omega^X) \wedge dh|_{x=x_0} \neq 0$,*where $\Lambda^{n-1}(i_X \Omega^X)$ is the vector space generated by $\{i_X(\omega_1) \wedge \cdots \wedge i_X \omega_{n-1}; \omega_1, \cdots, \omega_{n-1} \in \Omega^X\}$.*

In fact, this statement is a slight improvement of that of [HG1, p. 140].

A sketch of the proof is as follows. First, considering $X = \partial/\partial z_1$ and (\mathcal{L}) , it is straightforward to see that conditions (1) and (2) are satisfied. Moreover, they are coordinate-free; hence they are necessary. Second, let $(\omega_1, \cdots, \omega_{n-1})$ be a basis of Ω^X . By (2c), $i_X(\omega_j) = dz_{j+1}$, $j = 1, \cdots, n-1$ in some small enough neighbourhood of x_0 . Condition (2d) implies that $x \rightarrow z(x) = (z_1, \cdots, z_n)(x)$, with $z_1(x) = h(x)$, is a diffeomorphism. Using (2a) and (2b), it is easily seen that this transformation $z(x)$ sends (Σ) to (\mathcal{L}) .

Additionally, we can find in [HG1] an algorithm allowing us to construct the vector field X of Theorem 2 whenever it does exist. Therefore Theorem 2 is constructive.

Our goal in this paper is to prove the following theorem.

THEOREM 3. *Assume that (Σ) is given E.L.L.O.I. Then (i) if (Σ) is C^∞ and M is simply connected, then (Σ) is E.L.L.O.I.G.; (ii) if (Σ) is C^ω , then (Σ) is G.L.O.I.*

This result is, in fact, false in the multi-output case; it is very easy to construct C^ω -examples in which (Σ) is E.L.L.O.I. and not G.L.O.I. However, all these examples are E.L.L.O.I.G. We do not know, however, if (i) is true in the multi-output case.

Our method of proof uses the fact that, in the single output case, the structure-preserving pseudogroup is a particular Lie group. This, as will be shown in a later remark, is no longer true in the multi-output case. Moreover, we do not know if (i) is true in the nonsimply connected case, and, again, we have no counterexample. When

we prove these results, we will see that the step from the everywhere-local to the global case is equivalent to the trivialization of a certain cocycle with coefficients in a nonabelian group.

Our paper is organised as follows. In § 2 we state some preliminary results. Mainly, in the E.L.L.O.I. case, the vector field X of Theorem 2 is global, and the structure-preserving pseudogroup is a Lie group. In § 3 the proof of Theorem 3 is given. Section 4 is an appendix regrouping all the technical lemmas involved in the different proofs.

2. Some preliminary results. In this section, we will characterize all the vector fields X that match the conditions of Theorem 2, in terms of one of them. This will be a key point in determining the structure-preserving group.

Let us first fix some notation. We denote by $\omega_{u_k \dots u_1}^X$ the 2-form $L_{f_{u_k}}(i_X \omega_{u_{k-1} \dots u_1}^X) \wedge dh$ of Ω_k^X , $\omega_u^X = dL_{f_u}(h) \wedge dh \in \Omega_1^X$. Assume that (Σ) is L.L.O.I. at x^0 . The dimension of Ω^X is $n-1$, and a basis of Ω^X is of the form $(dz_2 \wedge dz_1, \dots, dz_n \wedge dz_1)$, where $z_1 = h$ and (z_1, \dots, z_n) is a coordinate system that linearizes (Σ) up to output injection.

PROPOSITION 4. *Assume that (Σ) is L.L.O.I. at x^0 and that (z_1, \dots, z_n) is as above. Then*

(a) *Every vector field Z matching the conditions of Theorem 2 is obtained as*

$$Z = \frac{\partial}{\partial z_1} + \sum_{k=2}^n c_k \frac{\partial}{\partial z_k} \quad \text{for some constant } c_2, \dots, c_n;$$

(b) *$(\tilde{z}_1, \dots, \tilde{z}_n)$ being another linearizing coordinate system for (Σ) , with $\tilde{z}_1 = h$, there is a unique triple (T, a, b) such that*

$$\begin{pmatrix} \tilde{z}_2 \\ \vdots \\ \tilde{z}_n \end{pmatrix} = T \begin{pmatrix} z_2 \\ \vdots \\ z_n \end{pmatrix} + z_1 a + b,$$

where T is a constant $(n-1) \times (n-1)$ invertible matrix and a and b are constant vectors in R^{n-1} .

Remark 5. Let A be the additive group $z_1 R^{n-1} \oplus R^{n-1}$ and G be the semidirect product group $GL(n-1, R) \times A$ (i.e., the group law is given by: if $(T, H) \in G$, $(T', H') \in G$, then $(T, H)(T', H') = (TT', H + TH')$).

If $E(x_0)$ is the set of all

$$\begin{pmatrix} z_2 \\ \vdots \\ z_n \end{pmatrix}$$

coordinate systems that linearize (Σ) up to output injection in some neighbourhood of x_0 , property (b) of Proposition 4 means that G acts freely and transitively on $E(x_0)$.

Remark 6. In the multi-output case, property (b) fails to be true even in the uncontrolled case. This is due to the presence of one-dimensional Brunowsky blocks in the linear observability canonical form. Consider the following uncontrolled system on R^3 :

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = 0, \quad \dot{x}_3 = 0, \quad y = (x_1, x_3),$$

and set $\hat{x}_1 = x_1$, $\hat{x}_2 = x_2 + \varphi(x_3)$, $\hat{x}_3 = x_3$; we get that

$$\dot{\hat{x}}_1 = \hat{x}_2 - \varphi(\hat{x}_3), \quad \dot{\hat{x}}_2 = 0, \quad \dot{\hat{x}}_3 = 0, \quad \hat{y} = (\hat{x}_1, \hat{x}_3) = y.$$

To obtain Proposition 4, we need the following technical lemmas, whose proof is postponed to the Appendix.

LEMMA 7 (see the Appendix for the proof). Let X, Z be vector fields meeting the conditions of Theorem 2. For any $u_1, \dots, u_k \in R^m$,

$$\begin{aligned} \omega_{u_k \dots u_1}^Z &= \omega_{u_k \dots u_1}^X + (i_X i_Z \omega_{u_1}^X) \omega_{u_k \dots u_2}^X \\ &\quad + \sum_{l=2}^{k-1} \left\{ (i_X i_Z \omega_{u_l \dots u_1}^X) \omega_{u_k \dots u_{l+1}}^X + \sum_{1 \leq i(1) < \dots < i(p+1-l) \leq p+1} P_{i(1) \dots i(k-l)} \omega_{u_{i(k-l)} \dots u_{i(1)}}^X \right\}, \end{aligned}$$

where $P_{i(1) \dots i(k-l)}$ belongs to the algebra generated by

$$\{i_X i_Z \omega_{u_{i(\tau)} \dots u_{i(1)}; u_{i(1)}, \dots, u_{i(\tau)}}^X \in \{u_1, \dots, u_k\}, 1 \leq \tau \leq l-1\}$$

and closed under the Lie derivative L_{f_u} , $u \in \{u_1, \dots, u_k\}$.

LEMMA 8 (see the Appendix for the proof). Let (Σ) be L.L.O.I. and let X and Z be vector fields meeting the conditions of Theorem 2. Then $i_X i_Z \omega_{u_k \dots u_1}^X$ is constant for any $u_1, \dots, u_k \in R^m$.

Proof of Proposition 4. (a) Let (z_1, \dots, z_n) be a linearizing coordinate system at z_0 , with $z_1 = h$. Choose a vector field X such that $L_X(z_1) = 1$ and $i_X(dz_j \wedge dz_1) = -dz_j$, $j = 2, \dots, n$. It is easy to see that X exists, is unique, and verifies conditions (1) and (2) of Theorem 2. Let Z be another vector field meeting the assumptions of Theorem 2. By Lemma 8, $i_Z i_X \omega_{u_k \dots u_1}^X$ is constant for any $u_1, \dots, u_k \in R^m$. In particular, $i_Z i_X(dz_j \wedge dz_1)$ are constant. However, $i_Z i_X(dz_j \wedge dz_1) = -i_Z(dz_j) = -L_Z(z_j)$, and

$$Z = \frac{\partial}{\partial z_1} + \sum_{j=2}^n L_Z(z_j) \frac{\partial}{\partial z_j}.$$

This implies part (a).

(b) Let $(T, H) \in G$, $T \in GL(n-1, R)$, $H = az_1 + b$, $a, b \in R^{n-1}$, with $h = z_1$. Let (z_1, \dots, z_n) be a coordinate system that linearizes (Σ) up to output injection. Set

$$\begin{pmatrix} \tilde{z}_2 \\ \vdots \\ \tilde{z}_n \end{pmatrix} = T \begin{pmatrix} z_2 \\ \vdots \\ z_n \end{pmatrix} + az_1 + b \quad \text{and} \quad \tilde{z}_1 = z_1.$$

Obviously, $(\tilde{z}_1, \dots, \tilde{z}_n)$ linearizes (Σ) up to output injection.

Conversely, z and \tilde{z} are two coordinate systems that linearize (Σ) up to output injection. Let X, Z be vector fields such that

- (i) $L_X(z_1) = L_Z(z_1) = 1$; and
- (ii) $i_X(dz_j \wedge dz_1) = -dz_j$, $j = 2, \dots, n$, $i_Z(d\tilde{z}_j \wedge dz_1) = -d\tilde{z}_j$, $j = 2, \dots, n$.

X and Z are unique and meet conditions of Theorem 2. By Lemmas 7 and 8,

$$\Omega^X = \bigoplus_{j=2}^n R dz_j \wedge dz_1 = \bigoplus_{j=2}^n R d\tilde{z}_j \wedge dz_1 = \Omega^Z.$$

There is then a constant matrix $T \in GL(n-1, R)$ such that

$$\begin{pmatrix} d\tilde{z}_2 \wedge dz_1 \\ \vdots \\ d\tilde{z}_n \wedge dz_1 \end{pmatrix} = T \begin{pmatrix} dz_2 \wedge dz_1 \\ \vdots \\ dz_n \wedge dz_1 \end{pmatrix}.$$

Applying i_Z to this last equality yields

$$\begin{pmatrix} d\tilde{z}_n \\ \vdots \\ d\tilde{z}_2 \end{pmatrix} = -T \begin{pmatrix} i_Z(dz_2 \wedge dz_1) \\ \vdots \\ i_Z(dz_n \wedge dz_1) \end{pmatrix}.$$

Using part (a) of Proposition 4, $Z = \partial/\partial z_1 + \sum_{l=2}^n a_l(\partial/\partial z_l)$; hence

$$\begin{pmatrix} d\tilde{z}_2 \\ \vdots \\ d\tilde{z}_n \end{pmatrix} = -T \begin{pmatrix} -dz_2 + a_2 dz_1 \\ \vdots \\ -dz_n + a_n dz_1 \end{pmatrix}.$$

This means that

$$\begin{pmatrix} \tilde{z}_2 \\ \vdots \\ \tilde{z}_n \end{pmatrix} = T \begin{pmatrix} z_2 \\ \vdots \\ z_n \end{pmatrix} + \begin{pmatrix} \tilde{a}_2 \\ \vdots \\ \tilde{a}_n \end{pmatrix} z_1 + \begin{pmatrix} \tilde{b}_2 \\ \vdots \\ \tilde{b}_n \end{pmatrix}.$$

Therefore G acts transitively on $E(x_0)$. It is obvious that the action is free.

The following lemma, showing the existence of a global vector field X meeting the assumptions of Theorem 2, is the key point for our developments.

LEMMA 9 (see the Appendix for the proof). *If (Σ) is E.L.L.O.I., there exists a global vector field X on M , which satisfies the conditions of Theorem 2.*

To finish the technicalities, it remains to state the following small lemma.

LEMMA 10 (see the Appendix for the proof). *Let M be a C^ω -connected manifold and let h be a C^ω function on M such that $dh(x) \neq 0$ for all $x \in M$. Let φ be a C^ω real-valued function on M satisfying $d\varphi \wedge dh = 0$. Then there exists an analytic function g on the image of h such that $\varphi = g \circ h$.*

3. Resolution of the G.L.O.I. problem. In this section, our aim is to give a geometric interpretation of the G.L.O.I. problem and to prove Theorem 3. First, let us recall some standard main notions that are needed for our purposes.

DEFINITION 11. Let (E, M, Y) be topological spaces. Let G be a group acting effectively on Y (that is, G can be identified with a subgroup of the group of homeomorphisms of Y). G is considered as a topological group together with the discrete topology.

Let $p: E \rightarrow M$ be a continuous projection. We say that (E, M, p, Y, G) is a fibre bundle with discrete structural group G and fibre Y if and only if there exists an open covering of M , $\{U_j\}_{j \in J}$ and a family of mappings $\{\Phi_j\}_{j \in J}$ such that

- (a) $\Phi_j: U_j \times Y \rightarrow p^{-1}(U_j)$ is a homeomorphism for every $j \in J$;
- (b) For all $j \in J$, and $(x, y) \in U_j \times Y$, $p(\Phi_j(x, y)) = x$. As a consequence,

$$\Phi_{j,x}: \begin{cases} Y \rightarrow p^{-1}(x), \\ y \rightarrow \Phi_j(x, y) \end{cases} \text{ is one-to-one and onto;}$$

- (c) For all $i, j \in J$ such that $U_{ij} = U_i \cap U_j \neq \emptyset$; for all $x \in U_{ij}$,

$$\Phi_{i,x}^{-1} \Phi_{j,x}: Y \rightarrow Y \text{ is an element of } G, \text{ and}$$

$$g_{ij}: \begin{cases} U_{ij} \rightarrow G, \\ x \rightarrow \Phi_{i,x}^{-1} \Phi_{j,x} \end{cases} \text{ is continuous.}$$

As soon as there is no ambiguity, we drop M, p and say that E is a fibre bundle with (structural discrete) group G and fibre Y . The set $\{g_{ij}\}$ is often called the “cocycle” associated with the fibre bundle E . We usually say that E is a “principal bundle” with (structural discrete) group G whenever the fibre Y is exactly G , acting on itself by left multiplication.

DEFINITION 12. With the same notation as in Definition 11, we say that a fibre bundle E with fibre Y and group G is trivial or M -isomorphic to the fibre bundle $M \times Y$ if and only if there exists a homeomorphism $\theta: M \times Y \rightarrow E$ such that

- (a) $\Phi_{j,x}^{-1} \circ \theta_x: Y \rightarrow Y$ belongs to G , where $\theta_x: Y \rightarrow p^{-1}(x)$, $y \rightarrow \theta_x(y) = \theta(x, y)$;

(b) The maps $g_j : U_j \rightarrow G$ are continuous,

$$x \rightarrow \Phi \circ_{j,x}^{-1} \theta_x.$$

Remark. Definition 12 is equivalent to the fact that the nonabelian cocycle $\{g_{ij}\}$ satisfies $g_{ij} = g_j g_i^{-1}$ for some $\{g_i\}_{i \in J}$; i.e., $\{g_{ij}\}$ is trivial in the sense that it is cohomologous to 1 (see [S]).

A cross section is a mapping $s : M \rightarrow E$, continuous, such that $p(s(x)) = x$, for all $x \in M$. Replacing everywhere above the continuity requirement by the C^r requirement, we say that E is a C^r bundle, s a C^r cross section, and so on.

THEOREM 13 (cf. [S, p. 36]). *A principal bundle E is trivial if and only if it admits a cross section.*

Assume now that (Σ) is E.L.L.O.I. Let $\{U_j\}_{j \in J}$ be an open covering of M such that the restriction $(\Sigma|_{U_j})$ (Σ restricted to U_j) is G.L.O.I. Let $\{z_1^{0j}, \dots, z_n^{0j}\}$, $j \in J$ be a privileged family of coordinate systems that linearize (Σ) up to output injection on each U_j . Recall (§ 2, Remark 5) that $E(x)$ was defined as the set of coordinate systems that linearize (Σ) up to output injection around x . Denote by E_x the set of germs at x of elements of $E(x)$.

Set $E = \bigcup_{x \in M} E_x$. Consider the map $p : E \rightarrow M$, which associates to z_x (germ at x of the linearizing coordinate system z) the point $p(z_x) = x$. We embed E with a topology: A basis of this topology is formed by the sets $\{z_x | x \in U\}$, where U is an open set of M , and z a linearizing coordinate system on U . This is nothing but the topology of the sheaf of germs of linearizing coordinate systems. E is a principal bundle with (structural discrete) group $G = GL(n-1, R) \times A$, and E is analytic or C^∞ , according to the fact that (Σ) is analytic or C^∞ . Part (b) of Proposition 4 means that G acts freely and transitively on E_x . Therefore we can identify G with each E_x .

Set

$$E_j = p^{-1}(U_j), \Phi_j^{-1}: \begin{cases} E_j \rightarrow U_j \times G, \\ z_x \rightarrow (x, g(z_x)), \end{cases}$$

where $g(z_x)$ is the unique element g of G such that $z_x^j = g^{-1} z_x^{0j}$. Part (b) of Proposition 4 gives the result $z_x^{0j} = g_{ij}(x) z_x^{0i}$ on U_{ij} ($U_{ij} = U_i \cap U_j$) and $g_{ij} : U_{ij} \rightarrow G$ is constant.

Now we prove the following intermediate result.

PROPOSITION 14. *If (Σ) is C^ω or (Σ) is C^∞ with simply connected underlying manifold M , then the fact that (Σ) is E.L.L.O.I. implies that E has a cross section.*

Proof. (Σ) being E.L.L.O.I., by Lemma 9 there is a global vector field X meeting the conditions of Theorem 2 on M .

Let $(\omega_{u_{i(1)} \dots u_{i(l)}}; u_{i(1)} \dots u_{i(l)} \in I)$ be a basis of Ω^X , where I is some index set whose construction is given in the Appendix. It holds that

$$(\omega_{u_{i(1)+1} \dots u_{i(l)}} = L_{f_{u_{i(1)+1}}} (i_X \omega_{u_{i(1)} \dots u_{i(l)}}) \wedge dh, \quad \omega_{u_{i(1)}} = dL_{f_{u_{i(1)}}} (h) \wedge dh).$$

To prove that the bundle E has a cross section, it is sufficient to show that

$$i_X \omega_{u_{i(1)} \dots u_{i(l)}} = dz_{u_{i(1)} \dots u_{i(l)}}.$$

In that case, the cross section will be given by

$$M \rightarrow E,$$

$$x \rightarrow s_x,$$

$s_x = (z_x^2, \dots, z_x^n)^T = z_x$, where $(z_2; \dots; z_n)$ is equal to $z_{u_{i(1)} \dots u_{i(1)}}, u_{i(1)} \dots u_{i(1)} \in I$ after reordering the index set I . Clearly, $i_X \omega_{u_{i(1)}} = -dL_{f_{u_{i(1)}}}(h) + L_X L_{f_{u_{i(1)}}}(h) dh$.

Condition (2c) of Theorem 2 expresses that $di_X \omega_{u_{i(1)}} = 0$; hence $dL_X L_{f_{u_{i(1)}}}(h) \wedge dh = 0$. Using Lemma 10, in the analytic case, there is a function $g_{u_{i(1)}}$ such that $L_X L_{f_{u_{i(1)}}}(h) = g_{u_{i(1)}} \circ h$. Let $G_{u_{i(1)}}$ be a primitive of $g_{u_{i(1)}}$; we obtain that $i_X(\omega_{u_{i(1)}} = dz_{u_{i(1)}}$, where $z_{u_{i(1)}} = -L_{f_{u_{i(1)}}}(h) + g_{u_{i(1)}} \circ h$.

In the C^∞ case, since M is simply connected, the same result holds for some $G_{u_{i(1)}}$.

Assume now that $i_X \omega_{u_{i(k)} \dots u_{i(1)}} = dz_{u_{i(k)} \dots u_{i(1)}}$;

$$\begin{aligned} \omega_{u_{i(k+1)} \dots u_{i(1)}} &= L_{f_{u_{i(k+1)}}}[i_X \omega_{u_{i(k)} \dots u_{i(1)}}] \wedge dh \\ &= L_{f_{u_{i(k+1)}}}[dz_{u_{i(k)} \dots u_{i(1)}}] \wedge dh \\ &= d[L_{f_{u_{i(k+1)}}}(z_{u_{i(k)} \dots u_{i(1)}})] \wedge dh; \end{aligned}$$

$$i_X \omega_{u_{i(k+1)} \dots u_{i(1)}} = -dL_{f_{u_{i(k+1)}}}(z_{u_{i(k)} \dots u_{i(1)}}) + L_X L_{f_{u_{i(k+1)}}}(z_{u_{i(k)} \dots u_{i(1)}}) dh.$$

Again, using $d(i_X \omega_{u_{i(k+1)} \dots u_{i(1)}}) = 0$, this implies that $dL_X L_{f_{u_{i(k+1)}}}(z_{u_{i(k+1)} \dots u_{i(1)}}) \wedge dh = 0$. The same argument as at the first step of the induction shows that

$$i_X \omega_{u_{i(k+1)} \dots u_{i(1)}} = dz_{u_{i(k+1)} \dots u_{i(1)}}.$$

The next proposition will give, with Proposition 14, part (i) of Theorem 3.

PROPOSITION 15. *In the C^∞ case, if E has a cross section, then (Σ) is E.L.L.O.I.G.*

Proof. Let $s: M \rightarrow E$ be a cross section. Denote by $s_x(x) \in R^{n-1}$ the valuation of some representative of s_x at x . Denote by τ the transformation $x \rightarrow (h(x), (s_x(x))^T)^T$ for some open subset V of R^n and where x^T denotes the transpose of x .

By using Proposition 4, τ sends $(\Sigma|_{U_j})$ the following system:

$$\mathcal{L}_j: \begin{cases} \dot{z}^j = A^j(u)z^j + \varphi^j(u, y), \\ y = z_1, \end{cases}$$

where

$$A^j(U) = \begin{Bmatrix} 0 \\ \vdots \\ \tilde{A}^j(U) \\ 0 \end{Bmatrix}$$

and $z^j(x) = \tau(x) \in V_j$, $V_j = \tau(U_j)$. $A^j(u)$ does not depend on x , and $A^j(u) = A^k(u)$ when $U_j \cap U_k \neq \emptyset$. Hence $A^j(u)$ does not depend on j , and we denote it by $A(u)$.

Writing $f_u(x) = f_u^1(x) + \varphi(u, x)$, where $f_u^1(x)$ is the pull-back by $\tau(x)$ of the vector field $A(u)z$, we get that (Σ) writes

$$\dot{x} = f_u^1(x) + \varphi(u, x), \quad y = h(x)$$

and τ locally transforms this system into the system \mathcal{L}_j ; i.e., the output injection $\varphi(u, x)$ is a global vector field and (Σ) is E.L.L.O.I.G.

To complete the proof of (ii) in Theorem 3, we need the following proposition.

PROPOSITION 16. *In the C^∞ case, (Σ) is G.L.O.I. if and only if E has a cross section.*

Necessity. Since (Σ) is G.L.O.I., there exists a diffeomorphism $\tau: M \rightarrow V$, V an open set of R^n with $\tau_1 = h$, and τ linearizes (Σ) up to output injection. Let s be the cross section $x \rightarrow s_x$, where s_x is the germ of $\tilde{\tau}$ at $x(\tilde{\tau} = (\tau_2, \dots, \tau_n)^T)$.

Sufficiency. The proof is the same as the one for Proposition 15, with the following remark: Denote by $A_k(u)$ the k th line of $A(u)$ and note that $d[L_{f_u}(\tau_k(x)) - A_k(u)\tau(x)] \wedge dh = 0$.

Using Lemma 10, $L_{f_u}(\tau_k(x)) - A_k(u)\tau(x)$ is a function $\varphi_k(u, h(x))$; hence

$$\dot{\tau}_k(x) = L_{f_u}(\tau_k(x)) = A_k(u)\tau(x) + \varphi_k(u, h(x)).$$

Therefore τ maps Σ to a system that is linear up to output injection (remember that $h(x) = \tau_1(x)$). τ is moreover injective: assume that $\tau(x^1) = \tau(x^2)$. Then x^1 and x^2 are indistinguishable, which contradicts the observability assumption.

Counterexample of the fact that, in the C^∞ case, (Σ) can be E.L.L.O.I. and not G.L.O.I. Let us recall that we do not know nonsimply connected examples for which (Σ) is E.L.L.O.I. and not E.L.L.O.I.G.

Set $M_1 = (]-2, 2[\times]-2, 2[) \setminus ([-1, 1] \times [-1, 1])$ and $M_2 = M_1 - (]1, 2[\times \{0\})$. Observe that M_1 is not simply connected and that M_2 is. Set

$$\varphi: \begin{cases} M_i \rightarrow R, \\ (x_1, x_2) \rightarrow \begin{cases} \exp(-1/(1/2 - x_1^2)) & \text{if } x_2 > 0 \text{ and } |x_1| < 1/\sqrt{2}, \\ 0 & \text{elsewhere.} \end{cases} \end{cases}$$

Consider (Σ_i)

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} 0 & u \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 0 \\ \varphi \end{pmatrix}; \quad i = 1, 2, \\ y = x_1.$$

(Σ_i) is observable in the sense of the rank condition; it is E.L.L.O.I. but not G.L.O.I.

Appendix.

Proof of Lemma 7. Let X, Z be vector fields meeting the conditions of Theorem 2. Let us first make the following remark.

Remark 17. Let $\omega = \sum_{j=2}^n a_j(x) dx_j \wedge dx_1$ and let X, Z be two vector fields such that $L_X(x_1) = L_Z(x_1) = 1$. Then

- (a) $i_Z \omega = i_X \omega + (i_X i_Z \omega) dx_1$,
- (b) $\omega = -(i_X \omega) \wedge dx_1 = -(i_Z \omega) \wedge dx_1$.

The proof of Lemma 7 is now an induction on k . The case where $k = 1$ follows from the fact that Ω_1^X does not depend on X : set $h = x_1$, $\omega_u^X = dL_{f_u}(x_1) \wedge dx_1 = \omega_u^Z$. Assume that the lemma is true for $1, \dots, k$. It holds that

$$\omega_{u_{k+1} \dots u_1}^Z = L_{f_{u_{k+1}}}(i_Z \omega_{u_k \dots u_1}^Z) \wedge dx_1.$$

By using the induction hypothesis,

$$\begin{aligned} i_Z \omega_{u_k \dots u_1}^Z &= i_Z \omega_{u_k \dots u_1}^X + (i_X i_Z \omega_{u_1}^X) i_Z \omega_{u_k \dots u_2}^X \\ &+ \sum_{l=2}^{k-1} \left\{ i_X i_Z \omega_{u_l \dots u_1}^X i_Z \omega_{u_k \dots u_{l+1}}^X + \sum_{1 \leq i(1) < \dots < i(k-l) \leq k} P_{i(1) \dots i(k-l)} i_Z \omega_{u_{i(k-l)} \dots u_{i(1)}}^X \right\}. \end{aligned}$$

By Remark 17(a),

$$\begin{aligned} i_Z \omega_{u_k \dots u_1}^Z &= i_X \omega_{u_k \dots u_1}^X + (i_X i_Z \omega_{u_k \dots u_1}^X) dx_1 \\ &+ (i_X i_Z \omega_{u_1}^X) i_X \omega_{u_k \dots u_2}^X + (i_X i_Z \omega_{u_1}^X) (i_X i_Z \omega_{u_k \dots u_2}^X) dx_1 \\ &+ \sum_{l=2}^{k-1} \left\{ (i_X i_Z \omega_{u_l \dots u_1}^X) i_X \omega_{u_k \dots u_{l+1}}^X + (i_X i_Z \omega_{u_l \dots u_1}^X) (i_X i_Z \omega_{u_k \dots u_{l+1}}^X) dx_1 \right. \\ &+ \sum_{1 \leq i(1) < \dots < i(k-l) \leq k} P_{i(1) \dots i(k-l)} i_X \omega_{u_{i(k-l)} \dots u_{i(1)}}^X \\ &\left. + P_{i(1) \dots i(k-l)} (i_X i_Z \omega_{u_{i(k-l)} \dots u_{i(1)}}^X) dx_1 \right\}. \end{aligned}$$

Applying Remark 17(b), we obtain, up to signs, that

$$\begin{aligned}
 \omega_{u_{k+1} \cdots u_1}^Z &= \omega_{u_{k+1} \cdots u_1}^X + (i_X i_Z \omega_{u_1}^X) \omega_{u_{k+1} \cdots u_2}^X \\
 &+ \sum_{l=2}^{k-1} \left\{ i_X i_Z \omega_{u_l \cdots u_1}^X \omega_{u_{k+1} \cdots u_{l+1}}^X + L_{f_{u_{k+1}}} (i_X i_Z \omega_{u_l \cdots u_1}^X) \omega_{u_k \cdots u_{l+1}}^X \right. \\
 &\quad + \sum_{1 \leq i(1) < \cdots < i(k-l) \leq k} P_{i(1) \cdots i(k-l)} \omega_{u_{k+1} \cdots u_{i(k-l)} \cdots u_{i(1)}}^X \\
 &\quad \left. + L_{f_{u_{k+1}}} (P_{i(1) \cdots i(k-l)}) \omega_{u_{i(k-l)} \cdots u_{i(1)}}^X \right\} \\
 &+ \left[i_X i_Z \omega_{u_k \cdots u_1}^X + (i_X i_Z \omega_{u_1}^X) (i_X i_Z \omega_{u_k \cdots u_2}^X) + \sum_{l=2}^{k-1} \{ (i_X i_Z \omega_{u_l \cdots u_1}^X) (i_X i_Z \omega_{u_k \cdots u_{l+1}}^X) \right. \\
 &\quad \left. + \sum_{1 \leq i(1) < \cdots < i(k-l) \leq k} P_{i(1) \cdots i(k-l)} i_X i_Z \omega_{u_{i(k-l)} \cdots u_{i(1)}}^X \} \right] \omega_{u_{k+1}}^X.
 \end{aligned}$$

Regrouping the terms of the same length, and using the fact that $P_{i(1) \cdots i(k-l)}$ is in the algebra spanned by

$$\{i_X i_Z \omega_{u_{i(\tau)} \cdots u_{i(1)}} \mid 1 \leq \tau \leq l-1; u_{i(1)}, \dots, u_{i(\tau)} \in \{u_1, \dots, u_k\}\}$$

and closed under the Lie derivative L_{f_u} , $u \in \{u_1, \dots, u_k\}$, we obtain Lemma 7 with k replaced by $k+1$.

Proof of Lemma 8. Let X, Z meet the assumptions of Theorem 2 at $x_0 \in M$. Let $\{\omega_{u_{i(l)} \cdots u_{i(1)}}^X; u_{i(l)} \cdots u_{i(1)} \in I, 1 \leq l \leq p\}$ be a basis of Ω^X obtained with the following procedure:

- consider $\{\omega_{u_{i(1)}}; u_{i(1)} \in I_1\}$, a basis of Ω_1^X (Ω_1^X is equal to $\text{span} \{dL_{f_u}(h) \wedge dh; u \in R^m\}$);
- complete this basis, to form a basis of $\Omega_1^X + \Omega_2^X$

$$\Omega_2^X = \text{Span}_R \{L_{f_u}(i_X \omega^X) \wedge dh; \omega^X \in \Omega_1^X, u \in R^m\}.$$

This new basis is $\{\omega_{u_{i(1)}}^X, u_{i(1)} \in I_1\} \cup \{\omega_{u_{i(2)} u_{i(1)}}^X, u_{i(2)} u_{i(1)} \in I_2\}$, where $\omega_{u_{i(2)} u_{i(1)}}^X = L_{f_{u_{i(2)}}}(i_X \omega_{u_{i(1)}}^X) \wedge dh$. Since, by assumption, $\dim_R \Omega^X = n-1$, this procedure ends.

Set $I = I_1 \cup \cdots \cup I_p$, $\text{Card}(I) = n-1$. The same procedure applied to the same index set I gives rise to $\{\omega_{u_{i(l)} \cdots u_{i(1)}}^Z; u_{i(l)} \cdots u_{i(1)} \in I\}$.

Remark 18. (i) (2d) of Theorem 2 states that $\{\omega_{u_{i(l)} \cdots u_{i(1)}}^X; u_{i(l)} \cdots u_{i(1)} \in I\}$ forms a basis of $\mathcal{E}_{x_0}^1 \wedge dh$, where $\mathcal{E}_{x_0}^1$ is the C^r module ($r = \infty$ or ω) of 1-forms in a neighbourhood of x_0 .

(ii) For any $k, 1 \leq k \leq p$ and for any v_1, \dots, v_k ,

$$\omega_{v_k \cdots v_1}^X = \sum_{l=1}^k \sum_{u_{i(l)} \cdots u_{i(1)} \in I_l} C_{u_{i(l)} \cdots u_{i(1)}}^{v_1 \cdots v_k} \omega_{u_{i(l)} \cdots u_{i(1)}}^X \quad \text{for some constant } C_{u_{i(l)} \cdots u_{i(1)}}^{v_1 \cdots v_k}.$$

Using Lemma 7,

$$\begin{aligned}
 \omega_{v_{p+1} \cdots v_1}^Z &= \omega_{v_{p+1} \cdots v_1}^X + (i_X i_Z \omega_{v_1}^X) \omega_{v_{p+1} \cdots v_2}^X \\
 &+ \sum_{l=2}^p \left\{ (i_X i_Z \omega_{v_l \cdots v_1}^X) \omega_{v_{p+1} \cdots v_{l+1}}^X \right. \\
 &\quad \left. + \sum_{1 \leq i(1) < \cdots < i(p+1-l) \leq p+1} P_{i(1) \cdots i(p+1-l)} \omega_{v_{i(p+1-l)} \cdots v_{i(1)}}^X \right\}.
 \end{aligned}$$

Let $v_{p+1} \cdots v_2 = u_{i(p)}^0 \cdots u_{i(1)}^0$ be a fixed element of I_p .

Using property (ii) of Remark 18,

$$\begin{aligned}
 (\mathcal{R}_1) \quad \omega_{u_{i(p)}^0 \cdots u_{i(1)}^0 v_1}^Z &= (C_{u_{i(p)}^0 \cdots u_{i(1)}^0 v_1} + (i_X i_Z \omega_{v_1}^X)) \omega_{u_{i(p)}^0 \cdots u_{i(1)}^0}^X \\
 &+ \sum_{u_{i(1)} \cdots u_{i(1)} \in I} \varphi_{u_{i(1)} \cdots u_{i(1)}}(x) \omega_{u_{i(1)} \cdots u_{i(1)}}^X, \\
 u_{i(l)} \cdots u_{i(1)} &\neq u_{i(p)}^0 \cdots u_{i(1)}^0.
 \end{aligned}$$

On the other hand,

$$\begin{aligned}
 \omega_{u_{i(p)}^0 \cdots u_{i(1)}^0 v_1}^Z &= \lambda_{u_{i(p)}^0 \cdots u_{i(1)}^0 v_1} \omega_{u_{i(p)}^0 \cdots u_{i(1)}^0}^Z + \sum_{u_{i(l)} \cdots u_{i(1)} \in I} \lambda_{u_{i(l)} \cdots u_{i(1)} v_1} \omega_{u_{i(l)} \cdots u_{i(1)}}^Z, \\
 u_{i(l)} \cdots u_{i(1)} &\neq u_{i(1)}^0 \cdots u_{i(1)}^0
 \end{aligned}$$

with constant $\lambda_{u_{i(l)} \cdots u_{i(1)} v_1}$ (by (2b) of Theorem 2).

Applying Lemma 7 to any term on the right-hand side of this formula gives

$$\begin{aligned}
 (\mathcal{R}'_1) \quad \omega_{u_{i(p)}^0 \cdots u_{i(1)}^0 v_1}^Z &= \lambda_{u_{i(p)}^0 \cdots u_{i(1)}^0 v_1} \omega_{u_{i(p)}^0 \cdots u_{i(1)}^0}^X \\
 &+ \sum_{u_{i(l)} \cdots u_{i(1)} \in I_l} \varphi_{u_{i(l)} \cdots u_{i(1)}}(x) \omega_{u_{i(l)} \cdots u_{i(1)}}^X, \\
 u_{i(l)} \cdots u_{i(1)} &\neq u_{i(1)}^0 \cdots u_{i(1)}^0.
 \end{aligned}$$

Using point (i) of Remark 18 and identifying (\mathcal{R}_1) and (\mathcal{R}'_1) , we get that $i_X i_Z \omega_{v_1}^X$ is the constant

$$\lambda_{u_{i(p)}^0 \cdots u_{i(1)}^0 v_1} - C_{u_{i(p)}^0 \cdots u_{i(1)}^0 v_1}.$$

We complete the proof by induction. Assume that $i_X i_Z \omega_{v_\tau \cdots v_1}^X$ are constant for any $v_1, \dots, v_\tau \in R^m$, $1 \leq \tau \leq k-1$. By Lemma 7,

$$\begin{aligned}
 \omega_{v_{k+p} \cdots v_1}^Z &= \omega_{v_{k+p} \cdots v_1}^X + i_X i_Z (\omega_{v_1}^X) \omega_{v_{k+p} \cdots v_2}^X \\
 &+ \sum_{l=2}^{k+p-1} \left\{ (i_X i_Z \omega_{v_l \cdots v_1}^X) \omega_{v_{k+p} \cdots v_{l+1}}^X \right. \\
 &\quad \left. + \sum_{1 \leq i(1) < \cdots < i(k+p-l) \leq k+p} P_{i(1) \cdots i(k+p-l)} \omega_{v_{i(k+p-l)} \cdots v_{i(1)}}^X \right\}.
 \end{aligned}$$

This induction hypothesis and the fact that $P_{i(k+p-l) \cdots i(1)}$ belongs to the algebra generated by

$$\{i_X i_Z \omega_{v_{i(\tau)} \cdots v_{i(1)}}^X, v_{i(1)}, \dots, v_{i(\tau)} \in \{v_1, \dots, v_{k+p}\}, 1 \leq \tau \leq l-1\}$$

and closed under L_{f_v} ; $v \in \{v_1, \dots, v_{k+p}\}$ imply that $P_{i(1) \cdots i(k+p-l)}$ are constant for $1 \leq l \leq k$.

Now, replacing $(v_{k+p}, \dots, v_{k+1})$ by $u_{i(p)}^0, \dots, u_{i(1)}^0$, a fixed element of I_p , and applying property (ii) of Remark 18, we get that

$$\begin{aligned}
 (\mathcal{R}_k) \quad \omega_{u_{i(p)}^0 \cdots u_{i(1)}^0 v_{k+p} \cdots v_1}^Z &= (C_{u_{i(p)}^0 \cdots u_{i(1)}^0 v_{k+p} \cdots v_1}^{v_1 \cdots v_k} + i_X i_Z \omega_{v_{k+p} \cdots v_1}^X) \omega_{u_{i(p)}^0 \cdots u_{i(1)}^0}^X \\
 &+ \sum_{u_{i(l)} \cdots u_{i(1)} \in I} \varphi_{u_{i(l)} \cdots u_{i(1)}}(x) \omega_{u_{i(l)} \cdots u_{i(1)}}^X \\
 &\text{for constant } C_{u_{i(p)}^0 \cdots u_{i(1)}^0}^{v_1 \cdots v_k}, \\
 u_{i(l)} \cdots u_{i(1)} &\neq u_{i(p)}^0 \cdots u_{i(1)}^0.
 \end{aligned}$$

Similarly to the case where $k=1$,

$$\begin{aligned}
 (\mathcal{R}'_k) \quad \omega_{u_{i(p)}^0 \cdots u_{i(1)}^0 v_{k+p} \cdots v_1}^Z &= \lambda_{u_{i(p)}^0 \cdots u_{i(1)}^0 v_{k+p} \cdots v_1}^{v_1 \cdots v_k} \omega_{u_{i(p)}^0 \cdots u_{i(1)}^0}^Z \\
 &+ \sum_{u_{i(l)} \cdots u_{i(1)} \in I} \lambda_{u_{i(l)} \cdots u_{i(1)} v_{k+p} \cdots v_1}^{v_1 \cdots v_k} \omega_{u_{i(l)} \cdots u_{i(1)}}^X \quad \text{with constant } \lambda, \\
 u_{i(l)} \cdots u_{i(1)} &\neq u_{i(p)}^0 \cdots u_{i(1)}^0.
 \end{aligned}$$

Lemma 7 applied to any 2-form on the right-hand side of formula (\mathcal{R}'_k) gives, after combination (\mathcal{R}_k) ,

$$i_X i_Z \omega_{v_k \cdots v_1}^X = \lambda_{u_{i(p)}^0 \cdots u_{i(1)}^0}^{v_1^1 \cdots v_k^k} - C_{u_{i(p)}^0 \cdots u_{i(1)}^0}^{v_1^1 \cdots v_k^k}.$$

Proof of Lemma 9. Assume that (Σ) is E.L.L.O.I. Let $\{U_j\}_{j \in J}$ be a family of open sets of M such that $(\Sigma_{U_j})[(\Sigma) \text{ restricted to } (U_j)]$ is G.L.O.I. Let $\{X^j\}_{j \in J}$ be the associated family of vector fields, X^j meeting the conditions of Theorem 2 on U_j .

The proof of Lemma 9 is based on the two following claims.

CLAIM 19. *There exists some index set: $I = I_1 U \cdots U I_p$, where*

(a) *For each $j \in J$, $\{\omega_{u_{i(1)} \cdots u_{i(l)}, u_{i(l)} \cdots u_{i(1)}}^j \in I\}$ form a basis of Ω^{X^j} (where, obviously, $\omega_{u_{i(1)} \cdots u_{i(l)}}^j = \omega_{u_{i(l)} \cdots u_{i(1)}}^{X^j} = L_{f_{u_{i(l)}}}(\omega_{u_{i(l-1)} \cdots u_{i(1)}}^{X^j})) \wedge dh$;*

(b) *$\{dL_{f_{u_{i(l)}}} \cdots L_{f_{u_{i(1)}}}(h) \wedge dh; u_{i(l)} \cdots u_{i(1)} \in I\}$ is a basis of $\mathcal{E}^1(M) \wedge dh$. $\mathcal{E}^1(M)$ is the $C^r(M)$ -module of 1-forms on M ($r = \infty$ or ω).*

CLAIM 20. (a) *It holds that*

$$\omega_{v_1}^j = dL_{f_{v_1}}(h) \wedge dh,$$

$$\omega_{v_2 v_1}^j = -dL_{f_{v_2}} L_{f_{v_1}}(h) \wedge dh + L_{X^j} L_{f_{v_1}}(h) dL_{f_{v_2}}(h) \wedge dh$$

and, for $k \geq 3$,

$$\begin{aligned} \omega_{v_k \cdots v_1}^j &= \pm dL_{f_{v_k}} \cdots L_{f_{v_1}}(h) \wedge dh \pm L_{X^j} L_{f_{v_1}}(h) dL_{f_{v_k}} \cdots L_{f_{v_2}}(h) \wedge dh \\ &\quad + \sum_{l=2}^{k-1} \left\{ \pm L_{X^j} L_{f_{v_l}} \cdots L_{f_{v_1}}(h) dL_{f_{v_k}} \cdots L_{f_{v_{l+1}}}(h) \wedge dh \right. \\ &\quad \left. + \sum_{1 \leq i(1) < \cdots < i(l) \leq k} P_{i(1) \cdots i(l)} dL_{f_{v_k}} \cdots \hat{L}_{f_{v_{i(l)}}} \cdots \hat{L}_{f_{v_{i(1)}}} \cdots L_{f_{v_1}}(h) \wedge dh \right\}, \end{aligned}$$

where $P_{i(1) \cdots i(l)}$ belongs to the algebra generated by

$$\{L_{X^j} L_{f_{u_{i(\tau)}}} \cdots L_{f_{u_{i(1)}}}(h); u_{i(\tau)}, \cdots, u_{i(1)} \in \{v_1, \cdots, v_k\}, 1 \leq \tau \leq l-1\}$$

and closed under L_{f_v} ; $v \in \{v_1, \cdots, v_k\}$, and where the symbol $\hat{}$ means "omission" of the corresponding term;

(b) *The equality $dL_{f_{v_k}} \cdots L_{f_{v_1}}(h) \wedge dh = \sum_{l=1}^k \psi_{u_{i(l)} \cdots u_{i(1)}}(x) dL_{f_{u_{i(l)}}} \cdots L_{f_{u_{i(1)}}}(h) \wedge dh$ holds;*

(c) *If $L_{X^j} L_{f_{v_{i(\tau)}}} \cdots L_{f_{v_{i(1)}}}(h) = L_{X^j} L_{f_{v_{i(\tau)}}} \cdots L_{f_{v_{i(1)}}}(h)$ on $U_{\zeta_j} = U_j \cap U_{\zeta} \neq \emptyset$, with $1 \leq \tau \leq l-1$, $v_{i(1)}, \cdots, v_{i(\tau)} \in \{v_1, \cdots, v_k\}$, then $P_{i(1) \cdots i(l)}$, associated to $\omega_{v_k \cdots v_1}^j$ in the above formula, coincides on U_{ζ_j} with $P_{i(1) \cdots i(l)}$ in the expression of $\omega_{v_k \cdots v_1}^{\zeta}$. This means that the algebraic expression of $P_{i(1) \cdots i(l)}$ on U_j does not depend explicitly on j .*

Proof of Claim 19. Let U_j, U_{ζ} be two open sets of the above family such that $U_{\zeta_j} = U_j \cap U_{\zeta} \neq \emptyset$. Let $B_j = \{\omega_{u_{i(l)} \cdots u_{i(1)}}^j, u_{i(l)} \cdots u_{i(1)} \in I\}$ be a basis of Ω^{X^j} . Using Claim 20(a) and the fact that B^j is a basis of $\mathcal{E}^1(U_j) \wedge dh$, where $\mathcal{E}^1(U_j)$ is the $C^r(U_j)$ -module of 1-forms on U_j , we get that this property is equivalent to

$$\{dL_{f_{u_{i(l)}}} \cdots L_{f_{u_{i(1)}}}(h) \wedge dh; u_{i(l)} \cdots u_{i(1)} \in I\},$$

which forms a basis of $\mathcal{E}^1(U_j) \wedge dh$.

Set $B_{\zeta} = \{\omega_{u_{i(l)} \cdots u_{i(1)}}^{\zeta}; u_{i(l)} \cdots u_{i(1)} \in I\}$, $\omega_{u_{i(l)} \cdots u_{i(1)}}^{\zeta} = L_{f_{u_{i(l)}}}(i_X \zeta(\omega_{u_{i(l-1)} \cdots u_{i(1)}}^{\zeta})) \wedge dh$ on U_{ζ} . An obvious induction shows that $\{\omega_{u_{i(l)} \cdots u_{i(1)}}^{\zeta}; u_{i(l)} \cdots u_{i(1)} \in I\}$ is a basis of $\Omega^{X_{\zeta}}$ restricted to $U_{j\zeta}$.

First, $\omega_u^{\zeta} = \omega_u^j$, since they do not depend explicitly on X . Assume that $\{\omega_{u_{i(k-1)} \cdots u_{i(1)}}^{\zeta}, u_{i(k-1)} \cdots u_{i(1)} \in I\}$ are C^r -independent:

$$\omega_{u_{i(k)} \cdots u_{i(1)}}^{\zeta} = \pm dL_{f_{u_{i(k)}}} \cdots L_{f_{u_{i(1)}}}(h) \wedge dh + A_{\zeta}^{\zeta},$$

$$\omega_{u_{i(k)} \cdots u_{i(1)}}^j = \pm dL_{f_{u_{i(k)}}} \cdots L_{f_{u_{i(1)}}}(h) \wedge dh + A^j,$$

where, by Claim 20(a), A^ζ and A^j are 2-forms in A_r , the $C^r(U_{\xi_j})$ -module generated by $\{dL_{f_{u_i(\tau)}} \cdots L_{f_{u_i(1)}}(h) \wedge dh, 1 \leq \tau \leq k-1, u_{i(\tau)} \cdots u_{i(1)} \in I\}$.

The $\omega_{u_i(k) \cdots u_i(1)}^j$ are R -independent by assumption (R -independent between them and with A_r). The observability condition (2d) of Theorem 2 implies that they are $C^r(U_{j\zeta})$ -independent (between them and with A_r). It follows that $\{\omega_{u_i(l) \cdots u_i(1)}^\zeta; u_{i(l)} \cdots u_{i(1)} \in I\}$ is a basis of Ω^{X^ζ} . Part (b) is an obvious consequence of (a) and Claim 20(a), (b).

Proof of Claim 20. For parts (a) and (b), see [HG1, pp. 141, 142, 144]. The algebraic form of $P_{i(1) \cdots i(l)}$ is obtained by induction, and it is easy to see that the formula does not depend on the U_j .

Having proved Claims 19 and 20, we can now argue the proof of Lemma 9. Note that $\{h\} \cup \{L_{f_{u_i(l)}} \cdots L_{f_{u_i(1)}}(h), u_{i(l)} \cdots u_{i(1)} \in I\}$ forms a coordinate system on each U_j , $j \in J$. It is sufficient to construct $\{\tilde{X}^j\}_{j \in J}$, a family of vector fields meeting the conditions of Theorem 2 and such that $L_{\tilde{X}^j} L_{f_{u_i(l)}} \cdots L_{f_{u_i(1)}}(h)$ coincides with $L_{\tilde{X}^\zeta} L_{f_{u_i(l)}} \cdots L_{f_{u_i(1)}}(h)$ on each $U_{j\zeta} = U_j \cap U_\zeta \neq \emptyset$. This will imply that \tilde{X}^j coincides with \tilde{X}^ζ on $U_{j\zeta}$, and thus there is a globally defined X on M . To construct $\{\tilde{X}^j\}_{j \in J}$, we return to the algorithm given in [HG1, pp. 144–147]. Let p be the integer subject to $I = I_1 U \cdots U I_p$. By Claim 20(a),

$$\begin{aligned} \omega_{v_{p+1} \cdots v_1}^j = & \pm dL_{f_{v_{p+1}}} \cdots L_{f_{v_1}}(h) \wedge dh \pm L_X L_{f_{v_1}}(h) dL_{f_{v_{p+1}}} \cdots L_{f_{v_2}}(h) \wedge dh \\ & + \sum_{l=2}^p \left\{ \pm L_X L_{f_{v_l}} \cdots L_{f_{v_1}}(h) dL_{f_{v_{p+1}}} \cdots L_{f_{v_{l+1}}}(h) \wedge dh \right. \\ & \left. + \sum_{1 \leq i(1) < \cdots < i(l) \leq p+1} P_{i(1) \cdots i(l)} dL_{f_{v_{p+1}}} \cdots \hat{L}_{f_{v_{i(l)}}} \cdots \hat{L}_{f_{v_{i(1)}}} \cdots L_{f_{v_1}}(h) \wedge dh \right\}. \end{aligned}$$

By replacing $v_{p+1} \cdots v_2$ by $u_{i(p)}^0 \cdots u_{i(1)}^0$, a fixed element of I_p , and by using part (b) of Claim 20, it follows that

$$\begin{aligned} \omega_{u_{i(p)}^0 \cdots u_{i(1)}^0}^j = & (\varphi_{u_{i(p)}^0 \cdots u_{i(1)}^0}^{j, v_1}(x) \pm L_X L_{f_{v_1}}(h)) dL_{f_{u_{i(1)}^0}} \cdots L_{f_{u_{i(1)}^0}}(h) \wedge dh \\ (\mathcal{F}_1) \quad & + \sum_{u_{i(l)} \cdots u_{i(1)} \in I} \varphi_{u_{i(l)} \cdots u_{i(1)}}^j(x) dL_{f_{u_{i(l)}}} \cdots L_{f_{u_{i(1)}}}(h) \wedge dh, \\ & u_{i(l)} \cdots u_{i(1)} \neq u_{i(p)}^0 \cdots u_{i(1)}^0. \end{aligned}$$

Expanding $\omega_{u_{i(p)}^0 \cdots u_{i(1)}^0}^j$ in the basis $\{\omega_{u_{i(l)} \cdots u_{i(1)}}^j, u_{i(l)} \cdots u_{i(1)} \in I\}$ gives

$$\begin{aligned} \omega_{u_{i(p)}^0 \cdots u_{i(1)}^0}^j = & C_{u_{i(p)}^0 \cdots u_{i(1)}^0}^{j, v_1} \omega_{u_{i(p)}^0 \cdots u_{i(1)}^0}^j \\ & + \sum_{u_{i(l)} \cdots u_{i(1)} \in I} C_{u_{i(l)} \cdots u_{i(1)}}^j \omega_{u_{i(l)} \cdots u_{i(1)}}^j, \\ & u_{i(l)} \cdots u_{i(1)} \neq u_{i(p)}^0 \cdots u_{i(1)}^0. \end{aligned}$$

Hence we get that

$$\begin{aligned} \omega_{u_{i(p)}^0 \cdots u_{i(1)}^0}^j = & C_{u_{i(p)}^0 \cdots u_{i(1)}^0}^{j, v_1} \omega_{u_{i(p)}^0 \cdots u_{i(1)}^0}^j \\ (\mathcal{F}_1') \quad & + \sum_{u_{i(l)} \cdots u_{i(1)} \in I} \tilde{\varphi}_{u_{i(l)} \cdots u_{i(1)} v_1}^j(x) dL_{f_{u_{i(l)}}} \cdots L_{f_{u_{i(1)}}}(h) \wedge dh, \\ & u_{i(l)} \cdots u_{i(1)} \neq u_{i(p)}^0 \cdots u_{i(1)}^0. \end{aligned}$$

The combination of (\mathcal{F}_1) and (\mathcal{F}_1') implies that $L_{X^j} L_{f_{v_1}}(h) = \theta_{v_1}(x) + \lambda_{v_1}^j$, where $\theta_{v_1}(x)$ does not depend on j and where $\lambda_{v_1}^j$ is a constant.

Recall that $\omega_{v_1}^j = dL_{f_{v_1}}(h) \wedge dh$ and that

$$\begin{aligned} i_{X^j} \omega_{v_1}^j = & -dL_{f_{v_1}}(h) + L_{X^j} L_{f_{v_1}}(h) dh \\ = & -dL_{f_{v_1}}(h) + \theta_{v_1}(x) dh + \lambda_{v_1}^j dh. \end{aligned}$$

Let $X^j(1)$ be the unique vector field such that

$$i_{X^j(1)}\omega_{v_1}^j = i_{X^j}\omega_{v_1}^j - \lambda_{v_1}^j dh = -dL_{f_{v_1}}(h) + \theta_{v_1}(x) dh,$$

$i_{X^j(1)} = i_{X^j}$ on some supplementary space of $\Omega_1^{X^j}$ in Ω^{X^j} , $L_{X^j(1)}(h) = 1$.

It is clear that $X^j(1)$ satisfies the assumptions of Theorem 2. Note that $L_{X^j(1)}L_{f_{v_1}}(h) = \theta_{v_1}(x)$, and, if $X^\xi(1)$ is defined on the same way as $X^j(1)$, with $U_{j\xi} \neq \emptyset$, then $L_{X^j(1)}L_{f_{v_1}}(h)$ coincides with $L_{X^\xi(1)}L_{f_{v_1}}(h)$ on $U_{j\xi}$.

Assume now that, for each $j \in J$, there exists $X^j(k)$, a vector field defined on U_j such that

(i) $X^j(k)$ meets the conditions of Theorem 2,

($\mathcal{P}(k)$) (ii) For all $j, \xi \in J$ such that $U_{j\xi} \neq \emptyset$, $L_{X^j(k)}L_{f_{v_\tau \dots v_1}}(h)$ coincides with

$$L_{X^\xi(k)}L_{f_{v_\tau}} \cdots L_{f_{v_1}}(h) \quad \text{on } U_{j\xi} \text{ for } 1 \leq \tau \leq k.$$

Denote by $\{\tilde{\omega}_{u_{i(l)} \dots u_{i(1)}}^j; u_{i(l)} \cdots u_{i(1)} \in I\}$ the basis of $\Omega^{X^j(k)}$ obtained in the usual manner. Consider

$$\begin{aligned} \tilde{\omega}_{v_{k+p+1} \dots v_1}^j &= \pm dL_{f_{v_{k+p+1}}} \cdots L_{f_{v_1}}(h) \wedge dh \pm L_{X^j(k)}L_{f_{v_1}}(h) dL_{f_{v_{k+p+1}}} \cdots L_{f_{v_2}}(h) \wedge dh \\ &+ \sum_{l=2}^{k+p} \left\{ \pm L_{X^j(k)}L_{f_{v_l}} \cdots L_{f_{v_1}}(h) dL_{v_{k+p+1}} \cdots L_{f_{v_{l+1}}}(h) \right. \\ &+ \sum_{1 \leq i(1) < \dots < i(l) \leq k+p+1} P_{i(1) \dots i(l)} \\ &\times dL_{f_{v_{k+p+1}}} \cdots \hat{L}_{f_{v_{i(l)}}} \cdots \hat{L}_{f_{v_{i(1)}}} \cdots L_{f_{v_1}}(h) \wedge dh \left. \right\}. \end{aligned}$$

Recall that $P_{i(1) \dots i(l)}$, together with its Lie derivatives, depends only on $L_{X^j}L_{f_{v_{i(\tau)}}} \cdots L_{f_{v_{i(1)}}}$ for $1 \leq \tau \leq l-1$. Using $\mathcal{P}(k)$ and Claim 20, we get that

$$\begin{aligned} \tilde{\omega}_{v_{k+p+1} \dots v_1}^j &= \sum_{u_{i(l)} \dots u_{i(1)} \in I} g_{u_{i(l)} \dots u_{i(1)}}(x) dL_{f_{u_{i(l)}}} \cdots L_{f_{u_{i(1)}}}(h) \wedge dh \\ &\pm L_{X^j(k)}L_{f_{v_{k+1}}} \cdots L_{f_{v_1}}(h) dL_{f_{v_{k+p+1}}} \cdots L_{f_{v_{k+2}}}(h) \wedge dh \\ &+ \sum_{l=k+2}^{k+p} \sum_{1 \leq i(1) < \dots < i(l) \leq k+p+1} P_{i(1) \dots i(l)} dL_{f_{v_{k+p+1}}} \cdots \\ &\times \hat{L}_{f_{i(l)}} \cdots \hat{L}_{f_{i(1)}} \cdots L_{f_{v_1}}(h) \wedge dh, \end{aligned}$$

where the coefficients $g_{u_{i(l)} \dots u_{i(1)}}$ associated respectively to $\tilde{\omega}_{v_{k+p+1} \dots v_1}^j$ and $\tilde{\omega}_{v_{k+p+1} \dots v_1}^\xi$ coincide on $U_{j\xi} \neq \emptyset$.

Now apply Claim 20(b) to $dL_{f_{v_{k+p+1}}} \cdots \hat{L}_{f_{v_{i(l)}}} \cdots \hat{L}_{f_{v_{i(1)}}} \cdots L_{f_{v_1}}(h) \wedge dh$, for $k+2 \leq 1 \leq k+p$, to get that

$$\begin{aligned} &\sum_{l=k+2}^{k+p} P_{i(1) \dots i(l)} dL_{f_{v_{k+p+1}}} \cdots \hat{L}_{f_{v_{i(l)}}} \cdots \hat{L}_{f_{v_{i(1)}}} \cdots L_{f_{v_1}}(h) \wedge dh \\ &= \sum_{l=1}^{p-1} \sum_{u_{i(l)} \dots u_{i(1)} \in I_l} g_{u_{i(l)} \dots u_{i(1)}}^j(x) dL_{f_{u_{i(l)}}} \cdots L_{f_{u_{i(1)}}}(h) \wedge dh. \end{aligned}$$

Replacing $v_{k+p+1} \cdots v_{k+2}$ by $u_{i(p)}^0 \cdots u_{i(1)}^0$ a fixed element of I , we obtain that

$$\begin{aligned} \mathcal{F}(k+1) &= (\tilde{g}_{u_{i(p)}^0 \dots u_{i(1)}^0}^j(x) \pm L_{X^j(k)}L_{f_{v_{k+1}}} \cdots L_{f_{v_1}}(h)) dL_{f_{u_{i(p)}^0}} \cdots L_{f_{u_{i(1)}^0}}(h) \wedge dh \\ &+ \sum_{u_{i(l)} \dots u_{i(1)} \neq u_{i(p)}^0 \dots u_{i(1)}^0} \tilde{g}_{u_{i(l)} \dots u_{i(1)}}^j(x) dL_{f_{u_{i(l)}}}(h) \wedge dh. \end{aligned}$$

As for $k = 1$,

$$\tilde{\omega}_{u_{i(1)}^0 \cdots u_{i(1)}^0 v_{k+1} \cdots v_1}^j = \sum_{ui(I) \cdots i(1) \in I} \lambda_{u_{i(1)}^0 \cdots u_{i(1)}^0}^{v_1 \cdots v_{k+1}} \tilde{\omega}_{u_{i(1)}^0 \cdots u_{i(1)}^0}^j,$$

where $\lambda_{u_{i(1)}^0 \cdots u_{i(1)}^0}^{v_1 \cdots v_{k+1}}$ are constants.

By using Claim 20(a), (b), this implies that

$$\begin{aligned} \mathcal{F}'_{(k+1)} \tilde{\omega}_{u_{i(1)}^0 \cdots u_{i(1)}^0 v_{k+1} \cdots v_1}^j &= \lambda_{u_{i(p)}^0 \cdots u_{i(1)}^0}^{v_1 \cdots v_{k+1}} dL_{f_{u_{i(p)}^0 \cdots u_{i(1)}^0}}(h) \wedge dh \\ &+ \sum_{ui(I) \cdots i(1) \neq u_{i(p)}^0 \cdots u_{i(1)}^0} g_{u_{i(1)}^0 \cdots u_{i(1)}^0}(x) dL_{f_{u_{i(1)}^0}} \cdots L_{f_{u_{i(1)}^0}}(h) \wedge dh. \end{aligned}$$

Combining $(\mathcal{F}_{(k+1)})$ and $(\mathcal{F}'_{(k+1)})$, we obtain that

$$L_{X^{j(k)}} L_{f_{v_{k+1}}} \cdots L_{f_{v_1}}(h) = \pm \tilde{g}_{u_{i(1)}^0 \cdots u_{i(1)}^0}(x) + \lambda_{u_{i(p)}^0 \cdots u_{i(1)}^0}^{v_1 \cdots v_{k+1}}.$$

Now we construct $X^j(k+1)$ as the unique vector field such that

$$\begin{aligned} i_{X^{j(k+1)}} \tilde{\omega}_{u_{i(k+1)}^0 \cdots u_{i(1)}^0}^j &= i_{X^{j(k)}} \tilde{\omega}_{u_{i(k+1)}^0 \cdots u_{i(1)}^0}^j - \lambda_{u_{i(p)}^0 \cdots u_{i(1)}^0}^{u_{i(k+1)}^0 \cdots u_{i(1)}^0} \quad \text{for } u_{i(k+1)}^0 \cdots u_{i(1)}^0 \in l_{k+1}; \\ i_{X^{j(k+1)}} &= i_{X^{j(k)}} \quad \text{otherwise;} \\ L_{X^{j(k+1)}}(h) &= 1. \end{aligned}$$

It is clear that $X^j(k+1)$ satisfies the conditions of Theorem 2.

Using $\mathcal{P}(k)$ and the fact that $g_{u_{i(p)}^0 \cdots u_{i(1)}^0}^{u_{i(k+1)}^0 \cdots u_{i(1)}^0}$ associated to $L_{X^{j(k)}} L_{f_{u_{i(k+1)}^0}} \cdots L_{f_{u_{i(1)}^0}}(h)$ and $L_{X^{\zeta(k)}} L_{f_{u_{i(k+1)}^0}} \cdots L_{f_{u_{i(1)}^0}}(h)$ coincide on $U_{J_{\zeta}} \neq \emptyset$, we obtain that

$$L_{X^{j(k+1)}} L_{f_{u_{i(k+1)}^0}} \cdots L_{f_{u_{i(1)}^0}}(h) = L_{X^{\zeta(k+1)}} L_{f_{u_{i(k+1)}^0}} \cdots L_{f_{u_{i(1)}^0}}(h) \quad \text{on } U_{J_{\zeta}} \neq \emptyset.$$

Let $\{\pi_{u_{i(1)}^0 \cdots u_{i(1)}^0}^j; u_{i(1)}^0 \cdots u_{i(1)}^0\}$ be a basis of $\Omega^{X^{j(k+1)}}$ defined as for $\Omega^{X^{j(k)}}$. Note that (i) $\pi_{u_{i(1)}^0 \cdots u_{i(1)}^0}^j = \tilde{\omega}_{u_{i(1)}^0 \cdots u_{i(1)}^0}^j$, for $1 \leq l \leq k$, and (ii)

$$\begin{aligned} \pi_{u_{i(l)}^0 \cdots u_{i(1)}^0}^j &= \pi_{u_{i(l)}^0 \cdots u_{i(1)}^0}^{\zeta}, \\ i_{X^{j(k+1)}}(\pi_{u_{i(l)}^0 \cdots u_{i(1)}^0}^j) &= i_{X^{\zeta(k+1)}}(\pi_{u_{i(l)}^0 \cdots u_{i(1)}^0}^{\zeta}) \end{aligned}$$

for all j, ζ subject to $U_{J_{\zeta}} \neq \emptyset$ and all $l, 1 \leq l \leq k+1$.

Part (ii) implies that $i_{X^{j(k+1)}} \pi_{v_{k+1} \cdots v_1}^j = i_{X^{\zeta(k+1)}} \pi_{v_{k+1} \cdots v_1}^{\zeta}$ on $\mathcal{U}_{i_{\zeta}} \neq \emptyset$ for all v_1, \dots, v_{k+1} , and then $L_{X^{j(k+1)}} L_{f_{v_{k+1}}} \cdots L_{f_{v_1}}(h) = L_{X^{\zeta(k+1)}} L_{f_{v_{k+1}}} \cdots L_{f_{v_1}}(h)$ on $\mathcal{U}_{i_{\zeta}} \neq \emptyset$.

This ends the proof of Lemma 9.

Proof of Lemma 10. First, note that since $dh(x) \neq 0$ for all x , the image of h is open. Second, the image by h of any open connected set is an (open) interval. The intersection of a finite number of images of such open connected sets is still an interval. Let $\{U_j\}_{j \in J}$ be an open covering of M by open connected sets U_j on which $\varphi = g^j \circ h$ for analytic g^j defined on $h(U_j)$. We set $h(y) = g^j(y)$ if $y \in h(U_j)$ and prove that the resulting map is well defined (this will imply analyticity).

Let x_1, x_N be given such that $x_1 \in U_1, x_N \in U_N, y = h(x_1) = h(x_N)$. By connectedness, pick a continuous path on M joining x_1 to x_N . By compactness, cover this path with a (finite) number of open sets $\{U_j\}_{j \in \tilde{J}}$. Reorder \tilde{J} so that $x_1 \in U_1, x_N \in U_N, U_j \cap U_{j+1} \neq \emptyset$, set $\mathcal{U}_i = U_1 \cup U_2 \cdots \cup U_i$. \mathcal{U}_i is connected. Set $\mathcal{I} = h(\mathcal{U}_i)$.

Assume (induction) that g is well defined on (the open interval) I_i . g^{i+1} is defined on $h(U_{i+1})$ and coincides with g on $h(\mathcal{U}_i \cap U_{i+1})$, but $h(\mathcal{U}_i \cap U_{i+1})$ is open in $h(\mathcal{U}_i) \cap h(U_{i+1})$, which is connected. By analyticity, g and g^{i+1} glue together on \mathcal{U}_{i+1} . It follows that $g^1(y) = g^N(y)$.

REFERENCES

- [B] W. M. BOOTHBY, *Global feedback linearization of locally linearizable systems*, Internal Report, Dept of Math., Washington University, St. Louis, MO, 1985.
- [BCC] G. BONNARD, N. COUENNE, AND F. CELLE, *Regularly persistent observers for bilinear systems*, in *New Trends in Nonlinear Control Theory*, Lecture Notes in Control and Inform. Sci., 122, Springer-Verlag, Berlin, New York, 1968, pp. 130–140.
- [BRG] D. BOSSANE, D. RAKOTOPARA, AND J. P. GAUTHIER, *Local and global immersion into linear systems up to output injection*, in *Proc. 28th IEEE Conf. on Decision and Control*, Tampa, FL, Dec. 13–15, 1989, pp. 2000–2004.
- [D] W. DAYAWANSA, *Geometry of the feedback linearization problem*, Ph.D. thesis, Sever Institute of Technology, Washington University, St. Louis, MO, August 1986.
- [FK] M. FLIESS AND I. KUPKA, *A finiteness criterion for nonlinear input-output differential systems*, *SIAM J. Control Optim.*, 21 (1983), pp. 721–728.
- [GCKS] J. P. GAUTHIER, F. CELLE, D. KASAKOS, AND G. SALLET, *Synthesis of nonlinear observers, a harmonic analysis approach*, *Math. Control Theory*, 22 (1989), pp. 291–322; *Math. Systems Theory*, to appear.
- [HG1] H. HAMMOURI AND J. P. GAUTHIER, *Bilinearization up to output injection*, *Systems Control Lett.*, 11 (1989), pp. 139–149.
- [HG2] ———, *The time varying linearization up to output injection*, in *Proc. 28th IEEE Conf. on Decision and Control*, Tampa, FL, Dec. 13–15, 1989, pp. 1038–1039.
- [K] A. J. KRENER, *The intrinsic geometry of dynamic observations*, in *Proc. of the Conference on the Algebraic and Geometric Methods in Nonlinear Control Theory*, June 3–5, 1985, Paris, CNRS, preprint.
- [KI] A. J. KRENER AND A. ISIDORI, *Linearization by output injection and nonlinear observers*, *Systems Control Lett.*, 3 (1983), pp. 47–52.
- [KR] A. J. KRENER AND W. RESPONDEK, *Nonlinear observers with linearizable error dynamics*, *SIAM J. Control Optim.*, 23 (1985), pp. 197–216.
- [LM] J. LEVINE AND R. MARINO, *Nonlinear systems immersion, observers and finite dimensional filters*, *Systems Control Lett.*, 7 (1986), pp. 133–142.
- [S] N. STEENROD, *The Topology of Fibre Bundles*, Princeton University Press, Princeton, NJ, 1951.

CONSTRAINED CONTROLLABILITY OF LINEAR DISCRETE NONSTATIONARY SYSTEMS IN BANACH SPACES*

VU NGOC PHAT†‡ AND TRINH CONG DIEU†

Abstract. This paper studies local null-controllability of linear infinite-dimensional, nonstationary, discrete-time systems of the form $x_{k+1} = A_k x_k + B_k u_k$, $u_k \in \Omega \subset U$, $x_k \in M_k \subset X$, where X, U are Banach spaces; A_k, B_k are linear bounded operators; M_k, Ω are given nonempty subsets. New necessary and sufficient conditions for local null-controllability are given. The main tool is the surjectivity theorem for convex multivalued mappings in Banach spaces.

Key words. linear discrete systems, constrained controllability, set-valued mappings, infinite-dimensional systems

AMS(MOS) subject classifications. 93B05, 93C15, 49B10

1. Introduction. Consider the following linear discrete-time control systems:

$$(1.1) \quad x_{k+1} = A_k x_k + B_k u_k, \quad k = 0, 1, \dots,$$

where $x_k \in M_k \subset X$, $u_k \in \Omega \subset U$, X, U are infinite-dimensional Banach spaces; M_k, Ω are given nonempty subsets; A_k, B_k are linear bounded operators.

A point $x \in X$ is said to be null-controllable at step N if there are admissible controls $u_k \in \Omega$, $k = 0, 1, \dots, N-1$ such that the corresponding solution x_k , $k = 0, 1, \dots, N$ of (1.1) satisfies $x_0 = x$, $x_k \in M_k$, $k = 1, 2, \dots, N-1$, $x_N = 0$.

Let S_N denote the set of all null-controllable at step N points of (1.1). Let

$$S = \bigcup_{N \geq 1} S_N.$$

System (1.1) is called locally null-controllable (locally null-controllable at step N , respectively) if and only if $0 \in \text{int } S$ ($0 \in \text{int } S_N$, respectively).

The study of controllability properties of linear dynamical systems has attracted much attention in the literature; refer to the surveys of Conti [1] and Faradzev, Phat, and Shapiro [3] for details. There are several results concerning the constrained controllability of linear discrete-time systems; see, for example, Murthy [4], Van Til and Shmitendorf [13], Phat [5], [8]. All of these works require the system either to be stationary or have unconstrained states, i.e., $M_k = X$. The problem of controllability of discrete-time systems with constraints on both control and state has received little attention. Some new conditions for local null-controllability and reachability of system (1.1) were obtained by Phat [6], [7] under the assumption that $0 \in \text{int } M_k$.

In this paper, to study local null-controllability of the control system (1.1) in the general case of constraints on both control and state, we apply some new results on convex multivalued analysis of Pshennichnyi [9] and Robinson [11] as surjectivity theorems and theorems on implicit functions for convex, closed, multivalued mappings in Banach spaces. This approach allows us to obtain general controllability tests in terms of support functionals in Banach spaces. Necessary and sufficient conditions for local null-controllability of control system (1.1) are established, where M_k is assumed to be an arbitrary convex closed set. The results of this paper can be considered an addendum to results of Phat [6], [7].

* Received by the editors July 17, 1989; accepted for publication (in revised form) June 26, 1991.

† Institute of Mathematics, P.O. Box 631, Bo Ho, Hanoi, Vietnam.

‡ This research was done while this author was visiting the Computer Centre of USSR Academy of Sciences, Moscow, Russia.

2. Notation and preliminary lemmas. We first adopt the following standard notation and definitions used throughout this paper. X^* , U^* denote the dual topological spaces of X , U . The closure, the interior, the relative interior, and the linear hull of a set M are denoted by \bar{M} , $\text{int } M$, $\text{ri } M$, and $\text{sp } M$, respectively. M^k denotes the set of all elements $x^k = (x_1, \dots, x_k)$ with $x_i \in M_i$, $i = 1, \dots, k$, and X^k represents a Banach space endowed with norm

$$\|x^k\| = \|x_1\| + \dots + \|x_k\|.$$

Let M^* be the dual cone to M at $0 \in M$ defined by

$$M^* = \{x^* \in X^*: \langle x^*, x \rangle \geq 0, x \in M\}.$$

To any multivalued mapping $T: X \Rightarrow U$, we associate, as in [9], the domain, the graph, and the adjoint map of T

$$\text{dom } T = \{x \in X: T(x) \neq \emptyset\},$$

$$\text{gf } T = \{(x, u) \in X \times U: u \in T(x)\},$$

$$T^*(u^*, z_0) = \{x^* \in X^*: (-x^*, u^*) \in (\text{gf } T - z_0)^*\},$$

where $z_0 \in \text{gf } T$.

A multivalued mapping $T: X \Rightarrow U$ is called convex (closed, respectively) if and only if $\text{gf } T$ is a convex (closed, respectively) subset of $X \times U$.

LEMMA 2.1 (extension of Robinson's surjectivity theorem). *Let $T: X \Rightarrow U$ be a convex, closed, multivalued mapping, whose range $T(X)$ is a set of the second category. Then, for every $x \in \text{dom } T$ and every neighbourhood B_x of x , we have that*

$$T(x) \cap \text{int } T(B_x) \neq \emptyset.$$

This result was proved by Duong and Tuy in [2].

LEMMA 2.2. *Let M be a convex subset of X . Assume that $0 \in M$. Then*

$$M^* = \{0\} \quad \text{and} \quad \text{int } M \neq \emptyset \Leftrightarrow 0 \in \text{int } M.$$

Proof. \Rightarrow : Assume that $0 \notin \text{int } M$. Then there is a nonzero functional $x^* \in X^*$ such that $\langle x^*, x \rangle \geq 0$ for all $x \in M$. This implies that $x^* \in M^*$; hence $M^* \neq \{0\}$.

\Leftarrow : Assume that $M^* \neq \{0\}$; i.e., there exists a nonzero functional $x^* \in X^*$ and $x^* \in M^*$. Therefore, by the definition of M^* , we have that $\langle x^*, x \rangle \geq 0$ for all $x \in M$. Since $0 \in \text{int } M$, we have that $x^* = 0$. \square

LEMMA 2.3 (theorem on implicit function). *Let $T: X \Rightarrow U$ be a convex multivalued mapping and $0 \in \text{gf } T$. Then*

$$0 \in \text{int dom } T \Leftrightarrow \text{int dom } T \neq \emptyset \quad \text{and} \quad T^*(0, 0) = \{0\}.$$

The proof of this lemma follows from Lemma 2.2 and from the theorem on implicit multivalued mappings proved by Pshennichnyi in [10]. \square

Let P be a linear bounded operator mapping from X into Y . Denote by P , P^{-1} the adjoint and inverse operators of P . It is clear that P^{-1} is a multivalued mapping defined by

$$P^{-1}y = \{x \in X: Px = y\}.$$

Let us set, for every $k = 1, 2, \dots$ and $i = 1, 2, \dots$,

$$P_{k,i} = A_k A_{k-1} \cdots A_i, \quad P_{k,k+1} = I, \quad P_{k,0} = -A_{k-1} \cdots A_1 A_0.$$

For every $u^k = (u_0, \dots, u_{k-1}) \in U^k$, we consider the following multivalued mapping $Q_k: U^k \Rightarrow X$:

$$(2.1) \quad Q_k u^k = P_{k,0}^{-1}(F_k u^k) \cap H_k u^k,$$

where

$$F_k u^k = \sum_{i=0}^{k-1} P_{k-1,i+1} B_i u_i,$$

$$H_k u^k = \bigcap_{i=1}^{k-1} G_i u^i,$$

$$G_k u^k = P_{k,0}^{-1}(F_k u^k - M_k).$$

Throughout the paper, unless otherwise specified, a control u^k , $k = 1, 2, \dots$, is called admissible if $u^k \in \Omega^k$ for $k = 1, 2, \dots$, and $Q_k(u^k) \neq \emptyset$. We assume that the sets M_k , $k = 1, 2, \dots$, are convex, closed, and $\text{int } M_k \neq \emptyset$, and Ω is a convex subset satisfying

$$\text{ri } \Omega \neq \emptyset, \quad 0 \in B_k \Omega, \quad k = 0, 1, \dots.$$

By definition, it is clear that the set of all null-controllable at step N points of system (1.1) is defined by the set $Q_N(\Omega^N)$. It also can be seen that the multivalued mapping $Q_k(\cdot)$ defined by (2.1) is convex, closed, and $0 \in \text{gf } Q_k$.

The following assumption plays a crucial role in what follows:

$$(2.2) \quad A_k^{-1}(B_k W) \cap \text{int } (-M_k) \cap \text{dom } P_{k,0}^{-1} \neq \emptyset, \quad k = 1, 2, \dots,$$

where $W = \overline{\text{sp}} \Omega$.

Note that, in the case where $M_k = X$ or $0 \in \text{int } M_k$, assumption (2.2) is immediately satisfied.

LEMMA 2.4. Assume that (2.2) holds true. Then

$$(2.3) \quad \text{int } H_k(W^k) \neq \emptyset, \quad k = 2, 3, \dots.$$

Proof. We first observe that if M_1, M_2 are convex subsets containing the origin and $\text{int } M_1 \neq \emptyset$, $\text{int } M_2 \cap M_1 \neq \emptyset$, then $\text{int } (M_1 \cap M_2) \neq \emptyset$. Using this fact, we now prove (2.3) by induction in $t = k-1, k-2, \dots, 1$. First, we show that

$$\text{int } (G_{k-1} W^{k-1} \cap G_{k-2} W^{k-2}) \neq \emptyset.$$

Indeed, in view of (2.2), there is a point $x_0 \in \text{int } (-M_{k-2})$ such that

$$A_{k-2} x_0 \in B_{k-2} W, \quad x_0 \in \text{dom } P_{k-2,0}^{-1}.$$

For any $y_0 \in P_{k-2,0}^{-1}(x_0)$, we have that

$$y_0 \in \text{int } G_{k-2} W^{k-2}.$$

Since

$$P_{k-1,0} y_0 = A_{k-2} P_{k-2,0} y_0 = A_{k-2} x_0,$$

we have that

$$P_{k-1,0} y_0 \in F_{k-1} W^{k-1} - M_{k-1},$$

which implies that

$$y_0 \in G_{k-1} W^{k-1}.$$

From the above remark, it follows that $\text{int } (G_{k-1} W^{k-1} \cap G_{k-2} W^{k-2}) \neq \emptyset$.

Now assume that (2.3) is satisfied for $t = 2, 3, \dots, k-1$; i.e.,

$$\text{int} \bigcap_{i=2}^{k-1} G_i W^i \neq \emptyset.$$

Let $x_0 \in A_1^{-1}(B_1 W) \cap \text{int}(-M_1)$. A point $y_0 \in P_{1,0}^{-1}(x_0)$ belongs to the interior of the set $P_{1,0}^{-1}(B_0 W - M_1)$. Since $A_1 x_0 \in B_1 W$, we obtain that

$$y_0 \in P_{i,0}^{-1}(F_i W^i - M_i), \quad i = 2, 3, \dots, k-1.$$

According to the above remark, we again obtain that

$$\text{int} \left\{ G_1 W^1 \cap \left(\bigcap_{i=2}^{k-1} G_i W^i \right) \right\} \neq \emptyset,$$

which completes the proof. \square

3. Controllability results. Let us now return to the control process (1.1). The proofs of our results on null-controllability theory are based on the surjectivity theorem and Lemma 2.3.

Let us define the following multivalued mapping:

$$T_N x = \{u^N \in \Omega^N : -P_{N,0} x + F_N u^N = 0, x \in H_N u^N\}.$$

It is obvious that T_N is a convex multivalued mapping from X into W^N and that

$$(3.1) \quad \text{dom } T_N = S_N = Q_N(\Omega^N).$$

THEOREM 3.1. Assume that (2.2) holds true. System (1.1) is locally null-controllable at step N if and only if

$$(3.2) \quad P_{N,0} X \subseteq \text{sp} \{B_{N-1} W, \dots, A_{N-1} \dots A_1 B_0 W\},$$

$$(3.3) \quad \{x^* \in X^*: x^* = A_0^* x_0^*, x_{k-1}^* = A_k^* x_k^* - m_k^*, m_k^* \in M_k^*, \\ k = 1, \dots, N-1, x_k^* \in (B_k \Omega)^*, k = 0, 1, \dots, N-1\} = \{0\}.$$

Proof. Necessity. Let system (1.1) be locally null-controllable at step N ; i.e., $0 \in \text{int } S_N$. By the definition of T_N and by (3.1), we have that

$$Q_N(W^N) = P_{N,0}^{-1}(F_N W^N) \cap H_N W^N.$$

Since

$$\text{int } Q_N(W^N) \neq \emptyset,$$

we have that

$$\text{int } P_{N,0}^{-1}(F_N W^N) \neq \emptyset,$$

which implies that

$$P_{N,0}^{-1}(F_N W^N) = X.$$

Thus we have that

$$P_{N,0} X \subseteq F_N W^N,$$

which proves (3.2). To prove (3.3), we set

$$\{x^* \in X^*: x^* = A_0^* x_0^*, x_{k-1}^* = A_k^* x_k^* - m_k^*, m_k^* \in M_k^*, k = 1, \dots, N-1, \\ x_k^* \in (B_k \Omega)^*, k = 0, \dots, N-1\} = R_N^*.$$

By Lemma 2.3, we have that $T_N^*(0, 0) = \{0\}$. Then it suffices to show that

$$(3.4) \quad R_N^* \subseteq T_N^*(0, 0).$$

For simplicity of formulation, we will prove (3.4) for $N = 2$.

Let $x^* \in R_2^*$. Then there are $x_0^* \in (B_0\Omega)^*$, $x_1^* \in (B_1\Omega)^*$, and $m_1^* \in M_1^*$ such that

$$x^* = A_0^* x_0^*, \quad x_0^* = A_1^* x_1^* - m_1^*.$$

By definition of T_2x , we have that

$$T_2x = \{u = \{u_0, u_1\} \in \Omega^2: A_1A_0x + A_2B_0u_0 + B_1u_1 = 0, A_0x + B_0u_0 \in M_1\}.$$

For every $(\bar{x}, \bar{u}) \in \text{gf } T_2$, we obtain that

$$\begin{aligned} \langle -x^*, \bar{x} \rangle + \langle 0, \bar{u} \rangle &= \langle -A_0^* A_1^* x_1^* + A_0^* m_1^*, \bar{x} \rangle \\ &= \langle x_1^*, -A_1 A_0 \bar{x} \rangle + \langle m_1^*, A_0 \bar{x} \rangle \\ &= \langle x_1^*, A_1 B_0 u_0 + B_1 u_1 \rangle + \langle m_1^*, A_0 \bar{x} \rangle. \end{aligned}$$

Since $x_0^* = A_1^* x_1^* - m_1^*$, we have that

$$\begin{aligned} \langle -x^*, \bar{x} \rangle &= \langle x_1^*, B_1 u_1 \rangle + \langle x_0^*, B_0 u_0 \rangle + \langle m_1^*, B_0 u_0 \rangle + \langle m_1^*, A_0 \bar{x} \rangle \\ &= \langle x_1^*, B_1 u_1 \rangle + \langle x_0, B_0 u_0 \rangle + \langle m_1^*, A_0 \bar{x} + B_0 u_0 \rangle. \end{aligned}$$

On the other hand, $x_i^* \in (B_i\Omega)^*$, $i = 1, 2$, and $A_0\bar{x} + B_0u_0 \in M_1^*$; then $\langle -x^*, \bar{x} \rangle \geq 0$. Therefore $x^* \in T_2^*(0, 0)$, which proves (3.4).

Sufficiency. Assume that conditions (3.2), (3.3) are satisfied. As we remarked above, $\text{dom } T_N = S_N = Q_N(\Omega^N)$. From (3.2), it follows that

$$P_{N,0}^{-1}(F_N W^N) = X.$$

By Lemma 2.4 and by the definition of $Q_N(W^N)$, we have that

$$\text{int } Q_N(W^N) \neq \emptyset.$$

Since $\text{ri } \Omega^N \neq \emptyset$, in view of Lemma 2.1 (i.e., the interior of Ω^N relative to the subspace $W^N = \overline{\text{sp}} \Omega^N$ is nonempty) we obtain that

$$\text{int } Q_N(\Omega^N) = \text{int dom } T_N \neq \emptyset.$$

Now we show that $T_N^*(0, 0) = \{0\}$, and then, from Lemma 2.3, it follows that $0 \in \text{int dom } T_N$; i.e., $0 \in \text{int } S_N$.

According to (3.3), it suffices to prove that $T_N^*(0, 0) \subseteq R_N^*$.

Let $N = 3$ (for every $N \geq 4$ the proof is analogous) and let $x^* \in T_3^*(0, 0)$. Then $(-x^*, 0) \in (\text{gf } T_3)^*$. From the definition of $T_3(x)$ and by an argument analogous to that used for the calculation of adjoint multivalued maps in [10], we can find that $v^* \in (F_3\Omega^3)^*$, $m^* \in (H_3\Omega^3)^*$ such that

$$x^* = -P_{3,0}^* v^* - m^* = A_0^* A_1^* A_2^* v^* - m^*.$$

According to a duality theorem for intersection of convex subsets in Banach spaces (see [9]), there exist functionals $\bar{m}_i^* \in (F_i\Omega^i - M_i)^*$, $i = 1, 2$ such that

$$m^* = -A_0^* A_1^* \bar{m}_2^* - A_0^* \bar{m}_1^*.$$

Since

$$-M_i \subseteq F_i\Omega^i - M_i, \quad i = 1, 2,$$

we have that

$$x^* = A_0^* A_1^* A_2^* v^* - A_0^* A_1^* m_2^* - A_0^* m_1^*,$$

where

$$m_i^* \in (-F_i \Omega^i + M_i)^* \subset M_i^*.$$

Setting

$$x_2^* = v^*, \quad x_1^* = A_2^* v^* - m_2^*, \quad x_0^* = A_1^* x_1^* - m_1^*,$$

we show that $x_i^* \in (B_i \Omega)^*$ for $i = 0, 1, 2$.

Indeed, by the definition of $F_3 \Omega^3$, we have that $B_2 \Omega \subset F_3 \Omega^3$. Then we have that $x_2^* = v^* \in (B_2 \Omega)^*$. Let $i = 1$. For every $(u_0, u_1) \in \Omega^2$, we obtain that

$$\begin{aligned} \langle x_1^*, A_1 B_0 u_0 + B_1 u_1 \rangle &= \langle A_2^* v^* - m_2^*, A_1 B_0 u_0 \rangle + \langle A_2^* v^* - m_2^*, B_1 u_1 \rangle \\ &= \langle v^*, A_2 A_1 B_0 u_0 \rangle + \langle v^*, A_2 B_1 u_1 \rangle + \langle -m_2^*, A_1 B_0 u_0 + B_1 u_1 \rangle. \end{aligned}$$

Since

$$F_2 \Omega^2 = A_1 B_0 \Omega + B_1 \Omega, \quad m_2^* \in (-F_2 \Omega^2 + M_2)^*,$$

we have that

$$-m_2^* \in (F_2 \Omega^2)^*.$$

On the other hand,

$$A_2 A_1 B_0 \Omega \subset F_3 \Omega^3, \quad A_2 B_1 \Omega \subset F_3 \Omega^3.$$

Hence

$$(3.6) \quad v^* \in (A_2 A_1 B_0 \Omega)^*, \quad v^* \in (A_2 B_1 \Omega)^*.$$

Combining (3.5) and (3.6), we have that $\langle x_1^*, A_1 B_0 u_0 + B_1 u_1 \rangle \geq 0$ for every $(u_0, u_1) \in \Omega^2$, which implies that $x_1^* \in (F_2 \Omega^2)^*$. Thus $x_1^* \in (B_1 \Omega)^*$.

Now let $i = 0$. For every $u_0 \in \Omega$, we have that

$$\langle x_0^*, B_0 u_0 \rangle = \langle A_1^* x_1^* - m_1^*, B_0 u_0 \rangle,$$

where $x_1^* \in (F_2 \Omega^2)^* \subset (B_1 \Omega)^*$, $m_1^* \in (-B_0 \Omega + M_1)^*$.

Since

$$-B_0 \Omega \subset M_1 - B_0 \Omega, \quad A_1 B_0 \Omega \subset F_2 \Omega^2,$$

we have that

$$-m_1^* \in (B_0 \Omega)^*, \quad x_1^* \in (A_1 B_0 \Omega)^*.$$

Hence, for every $u_0 \in \Omega$,

$$\langle x_0^*, B_0 u_0 \rangle = \langle x_1^*, A_1 B_0 u_0 \rangle + \langle -m_1^*, B_0 u_0 \rangle \geq 0,$$

which implies that $x_0^* \in (B_0 \Omega)^*$. The proof is complete. \square

THEOREM 3.2. *System (1.1) is locally null-controllable if and only if it is locally null-controllable at timestep $N \geq 1$.*

Proof. Let us consider the multivalued mappings $Q_N : W^N \rightrightarrows X$ defined by (2.1). Then, as we remarked above, $S_N = Q_N(\Omega^N)$. Let

$$S \cup S_N, \quad n \geq 1$$

The local null-controllability of system (1.1) implies that $0 \in \text{int } S$. By the Baire category theorem, there is a number $N \geq 1$ such that the set S_N is of the second category. Since the interior of Ω^N (relative to subspace W^N) is nonempty, using Lemma 2.1, we obtain that

$$\text{int } Q_N(\Omega^N) \neq \emptyset.$$

On the other hand, observe that

$$S_k \subseteq S_{k+1} \quad \text{for } k = 1, 2, \dots.$$

Hence, by the same arguments as in Lemma 1 [12], we can find a number $N_1 \geq N$ such that $0 \in \text{int } S_{N_1}$, which means that system (1.1) is locally null-controllable at step N_1 . The sufficiency is obvious. The proof is complete. \square

Combining Theorems 3.1 and 3.2, we obtain the following result.

THEOREM 3.3. *Assume that (2.2) holds true. System (1.1) is locally null-controllable if and only if (3.2), (3.3) hold true for some $N \geq 1$.*

Example 3.1. Consider the following system in l_2 :

$$(3.7) \quad \begin{aligned} x_{(k+1)} &= A_{(k)}x_{(k)} + u_{(k)}, & k &= 0, 1, \dots, N-1, \\ u_{(k)} &\in \Omega \subset l_2, & x_{(k)} &\in M_k \subset l_2, \end{aligned}$$

where

$$\begin{aligned} A_{(0)} &: (x_1, x_2, \dots) \rightarrow (x_1, x_3, x_5, \dots), \\ A_{(k)} &: (x_1, x_2, \dots) \rightarrow \frac{1}{k} (x_{k+1}, x_{k+2}, \dots), \\ \Omega &= \{u = (u_1, u_2, \dots) \in l_2 : u_i = 0, i = 1, \dots, N, \|u\| \leq 1\}, \\ M_k &= \{(x_1, x_2, \dots) \in l_2 : x_{N+k+1} \geq 0\}. \end{aligned}$$

It is easy to see that $\text{int } \Omega = \emptyset$, but $\text{ri } \Omega \neq \emptyset$ and

$$\begin{aligned} A_{(0)}^* &: (x_1, x_2, \dots) \rightarrow (x_1, 0, x_2, 0, x_3, 0, \dots), \\ A_{(k)}^* &: (x_1, x_2, \dots) \rightarrow \frac{1}{k} (0, \dots, 0, x_{k+1}, x_{k+2}, \dots), \\ \Omega^* &= \{u = (u_1, u_2, \dots) \in l_2 : u_{N+1} = u_{N+2} = \dots = 0\}, \end{aligned}$$

$$M_k^* = \{(x_1, x_2, \dots) \in l_2 : x_{N+k+1} \geq 0, x_i = 0, i \neq N+k+1\}.$$

In this case, conditions (3.2), (3.3) are satisfied. To verify (2.2), we take a number $y_{N+k+1} < 0$ for $k \in \{1, 2, \dots, N-1\}$ and set

$$x_{(k)} = (0, \dots, 0, y_{N+k+1}, 0, \dots).$$

We have that $x_{(k)} \in \text{int } (-M_k)$ and

$$A_{(k)}x_{(k)} = (0, \dots, 0, y_{N+1}, 0, \dots).$$

Hence

$$x_{(k)} \in A_{(k)}^{-1}(B_{(k)}W).$$

On the other hand, we have that

$$x_{(k)} \in \text{dom } P_{k,0}^{-1}.$$

Therefore (2.2) holds true, and system (3.7) is locally null-controllable at step N .

Acknowledgments. The authors thank the referees for many helpful comments and remarks.

REFERENCES

- [1] R. CONTI, *Linear controllability in finite-dimension*, Matematiche (Catania), 36 (1986).
- [2] P. C. DUONG AND H. TUY, *Stability, surjectivity and local invertibility of nondifferential mapping*, Acta Math. Vietnam., 1 (1978), pp. 89-105.
- [3] R. G. FARADZEV, V. N. PHAT, AND A. SHAPIRO, *Controllability theory of discrete dynamical systems*, Avtomat. i Telemekh., 1 (1986), pp. 5-24; English translation, Automat. Remote Control, 47 (1986), pp. 1-20.
- [4] D. MURTHY, *Controllability of a linear positive dynamic system*, Internat. J. Systems Sci., 17 (1986), pp. 49-54.
- [5] V. N. PHAT, *Controllability of discrete-time systems with nonconvex constrained controls*, Optimization, 3 (1983), pp. 371-375.
- [6] ———, *Controllability of nonlinear discrete-times systems without differentiability assumption*, Optimization, 1 (1988), pp. 133-142.
- [7] ———, *Controllability of linear nonstationary discrete systems with constrained states*, Avtomat. i Telemekh., 8 (1988), pp. 51-59; English translation, Automat. Remote Control, 49 (1988), pp. 998-1004.
- [8] ———, *Controllability of linear discrete systems with multiple delays on controls and states*, Internat. J. Control, 5 (1989), pp. 1645-1654.
- [9] B. N. PSHENNICHNYI, *Necessary Conditions of Extremums*, Nauka, Moscow, 1982.
- [10] ———, *Theorems on implicit functions for multivalued mappings*, Kibernetika, 4 (1986), pp. 35-43.
- [11] S. M. ROBINSON, *Regularity and stability for convex multi-valued functions*, Math. Oper. Res., 11 (1976), pp. 130-143.
- [12] N. K. SON, *Controllability of linear discrete systems with constrained controls in Banach spaces*, Control Cybernet., 1-2 (1981), pp. 5-17.
- [13] R. VAN TIL AND W. E. SHMITENDORF, *Constrained controllability of discrete-times systems*, Internat. J. Control, 43 (1986), pp. 941-954.

EXTENDED QUADRATIC CONTROLLER NORMAL FORM AND DYNAMIC STATE FEEDBACK LINEARIZATION OF NONLINEAR SYSTEMS*

WEI KANG[†] AND ARTHUR J. KRENER[†]

Abstract. In this paper, a set of extended quadratic controller normal forms of linearly controllable nonlinear systems is given, which is the generalization of the Brunovsky form of linear systems. A set of invariants under the quadratic changes of coordinates and feedbacks is found. It is then proved that any linearly controllable nonlinear system is linearizable to second degree by a dynamic state feedback.

Key words. nonlinear systems, quadratic normal forms, invariants, dynamic state feedbacks

AMS(MOS) subject classifications. 93C10, 93C15

1. Introduction. It is well known that there are four normal forms of linear systems: controllable, observable, controller, and observer form. The nonlinear generalizations of these four linear normal forms were given and discussed in Krener [12], Hunt and Su [5], Jakubczyk and Respondek [8], Brockett [1], and Sommer [16], among others. For a system in controller normal form, the design of a stabilizing state feedback control law is a straightforward task. Unfortunately, most controllable systems do not admit a controller normal form, and even when one does, the transformation of a system into controller normal form involves solving a system of first-order partial differential equations (PDEs), which numerically can be quite difficult. For these reasons, the approximate versions of nonlinear controller and observer normal forms were introduced in Krener [11], Krener et al. [13], Phelps and Krener [14], and Karahan [10], among others. It was proved that for certain kinds of nonlinear controllable systems, we can find a nonlinear change of coordinates and nonlinear state feedback that transforms the system into the linear approximation of the plant dynamics, which is accurate to second or higher degree. The computation of such a change of coordinates and state feedback is reduced to solving a set of linear equations. However, these linear equations are not always solvable, and most of the nonlinear systems do not admit such a linear approximation.

In this paper, a set of extended quadratic controller normal forms of linearly controllable systems with single input is given (Theorems 2 and 3). We can consider these normal forms as the extension of the Brunovsky form to the nonlinear systems. Then we prove that, given a nonlinear system, there exists a dynamic state feedback so that the extended system has a linear approximation that is accurate to at least second degree (Theorem 4). This means that any linearly controllable nonlinear system is linearizable to second degree by a dynamic state feedback (see the corollaries).

In this paper, we only consider the single-input systems. The generalization to multi-input systems will be given in another paper.

2. Extended quadratic controller form and dynamic state feedback linearization. From Brunovsky [2] (see also Kailath [9]), we know that any controllable linear system can be transformed into a controller form by a linear change of coordinates. If, in addition, we also allow linear change of coordinates in the input space and linear state

* Received by the editors November 19, 1990; accepted for publication (in revised form) August 2, 1991. This research was supported in part by Air Force Office of Scientific Research grant AFOSR 91-0228.

[†] Department of Mathematics, Institute of Theoretical Dynamics, University of California, Davis, California 95616.

feedback, any controllable linear system can be transformed into a Brunovsky form. Under the linear change of coordinate and state feedback, the Brunovsky form is a normal form for controllable linear systems. This result is summarized in the following theorem. The change of coordinates and state feedback used in this theorem is

$$(2.1) \quad \xi = Tx, \quad v = \alpha x + \beta \mu,$$

where T is a constant $n \times n$ nonsingular matrix, α is a row vector, and β is a nonzero real number.

THEOREM 1. *Consider a single-input, time-invariant linear system*

$$(2.2) \quad \dot{\xi} = F\xi + G\mu.$$

If it is controllable, then, by a suitable change of coordinates and state feedback (2.1), this linear system can be transformed into the following system of Brunovsky form:

$$(2.3) \quad \dot{x} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ 1 \end{bmatrix} v.$$

In the following, we give a nonlinear generalization of Brunovsky form from the quadratic approximation point of view. We study the following nonlinear systems:

$$(2.4) \quad \dot{\xi} = f(\xi) + g(\xi)\mu,$$

where $f(\xi)$ and $g(\xi)$ are nonlinear vector fields such that

$$(2.5) \quad f(0) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Throughout this paper, we use the following notation:

$$(2.6a) \quad A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}_{n \times n}, \quad B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}_{n \times 1};$$

$$(2.6b) \quad f^{[2]}(\xi) = \begin{bmatrix} f_1^{[2]}(\xi) \\ f_2^{[2]}(\xi) \\ \vdots \\ f_n^{[2]}(\xi) \end{bmatrix}, \quad g^{[1]}(\xi) = \begin{bmatrix} g_1^{[1]}(\xi) \\ g_2^{[1]}(\xi) \\ \vdots \\ g_n^{[1]}(\xi) \end{bmatrix};$$

$$(2.6c) \quad F = \frac{\partial f}{\partial \xi}(0), \quad G = g(0).$$

The superscripts of $f^{[2]}(\xi)$ and $g^{[1]}(\xi)$ denote that $f_i^{[2]}(\xi)$ and $g_i^{[1]}(\xi)$ are homogeneous polynomials of second and first degree in ξ . This kind of superscript will also be applied to some other vector fields and functions (e.g., $\alpha^{[2]}(x)$ or $\beta^{[1]}(x)$).

DEFINITION 1. If (F, G) , the linear part of system (2.4), is controllable, we call (2.4) a linearly controllable system. In this paper, we always assume that a nonlinear system is linearly controllable.

As mentioned in § 1, most linearly controllable nonlinear systems do not admit a controller normal form; therefore they cannot be transformed into the Brunovsky form (2.3). Theorem 1 implies that there exists a linear change of coordinates and linear state feedback (2.1), which transforms a nonlinear system (2.4) into the following system:

$$(2.7) \quad \dot{x} = Ax + Bv + f^{[2]}(x) + g^{[1]}(x)v + O(x, v)^3,$$

where (A, B) has Brunovsky form (2.6a). So Brunovsky form is a normal form for the linear part of linearly controllable nonlinear system (2.4). The question is: What is a normal form of the quadratic terms in this system? We answer this in the next theorem.

To solve this problem, let us consider the following nonlinear systems, the linear part of which is in Brunovsky form:

$$(2.8) \quad \dot{\xi} = A\xi + B\mu + f^{[2]}(\xi) + g^{[1]}(\xi)\mu + O(\xi, \mu)^3.$$

Since the linear part of (2.8) is already in Brunovsky form, we want to leave it invariant under a change of coordinates and state feedback. Therefore we consider the change of coordinates and state feedback of the following form:

$$(2.9) \quad \xi = x + \phi^{[2]}(x), \quad v = \mu + \alpha^{[2]}(x) + \beta^{[1]}(x)\mu.$$

Here $\phi^{[2]}(x)$ is an n -dimensional vector field whose entries are homogeneous polynomials of second degree in x , $\alpha^{[2]}(x)$ is a homogeneous polynomial of second degree, and $\beta^{[1]}(x)$ is a polynomial of first degree. The transformation given by (2.9) has two virtues. The first is that it leaves invariant the linear part of (2.8). The second is that the nonlinear coordinates ξ and x agree to the first degree.

Now we can define the normal form of the quadratic terms in (2.8) under the change of coordinates and state feedback (2.9).

THEOREM 2. *By a change of coordinates and state feedback (2.9), system (2.8) can be transformed into one and only one of the following systems:*

$$(2.10a) \quad \dot{x} = Ax + Bv + \tilde{f}^{[2]}(x) + O(x, v)^3,$$

where

$$(2.10b) \quad \tilde{f}^{[2]}(x) = \begin{bmatrix} \tilde{f}_1^{[2]}(x) \\ \tilde{f}_2^{[2]}(x) \\ \vdots \\ \tilde{f}_n^{[2]}(x) \end{bmatrix},$$

$$(2.10c) \quad \tilde{f}_i^{[2]}(x) = \begin{cases} \sum_{j=i+2}^n a_{ij}x_j^2 & 1 \leq i \leq n-2, \\ 0 & i = n-1 \text{ or } n. \end{cases}$$

DEFINITION 2. A system such as (2.10) is said to be in extended quadratic controller form.

An alternate set of normal forms is given in the next theorem, where $\tilde{f}^{[2]}(x)$ is zero.

THEOREM 3. *By a change of coordinates and state feedback (2.9), system (2.8) can be transformed into one and only one of the following systems:*

$$(2.11a) \quad \dot{x} = Ax + Bv + \tilde{g}^{[1]}(x)v + O(x, v)^3,$$

where

$$(2.11b) \quad \tilde{g}^{[1]}(x) = \begin{bmatrix} \tilde{g}_1^{[1]}(x) \\ \tilde{g}_2^{[1]}(x) \\ \vdots \\ \tilde{g}_n^{[1]}(x) \end{bmatrix},$$

$$(2.11c) \quad \tilde{g}_i^{[1]}(x) = \begin{cases} 0 & i = 1 \text{ or } n, \\ \sum_{j=n-i+2}^n a_{ij}x_j & \text{others.} \end{cases}$$

Remark 1. If both the linear and quadratic changes of coordinates and state feedbacks are used, what is the normal form of a nonlinear control system such as (2.4) under this larger transformation group? In fact, all the linear changes of coordinates and state feedbacks that leave the Brunovsky form invariant are $z = cx$, $\tilde{v} = c^{-1}v$, where c is a constant. When we apply this linear transformation to the normal form (2.10), the resulting quadratic part is $c^{-1}\tilde{f}^{[2]}(x)$. Let P denote the projective space induced by the linear space

$$\{\tilde{f}^{[2]}(x); \tilde{f}^{[2]}(x) \text{ is in the normal form (2.10)}\}.$$

The above linear transformation does not change $\tilde{f}^{[2]}(x)$ in the projective space P . Therefore, under both the linear and quadratic transformations, the family of the normal forms of systems such as (2.4) is a projective space plus the origin $\tilde{f}^{[2]}(x) = 0$.

Since most nonlinear systems (2.4) do not admit a controller form, they cannot be completely linearized by a change of coordinates and a state feedback. We wish to use (2.9) to transform (2.4) into a linear system plus an error of second or higher degree, below:

$$(2.12) \quad \dot{x} = Fx + Gv + O(x, v)^3.$$

A system with this property is said to be quadratically linearizable by (2.9). From the result of Theorem 2, we know that system (2.8) is quadratically linearizable by (2.9) if and only if the corresponding extended quadratic controller form (2.10) satisfies

$$(2.13) \quad \tilde{f}^{[2]}(x) = 0.$$

Therefore most nonlinear systems are not quadratically linearizable by state feedback. In the following, we introduce a method of linearizing a nonlinear system to the second degree by a dynamic state feedback. The concept of dynamic state feedback was introduced and studied in Singh [15] and Charlet, Lévine, and Marino [3], [4].

DEFINITION 3. A dynamic state feedback is a system

$$(2.14) \quad \begin{aligned} \dot{\omega} &= a(\xi, \omega) + b(\xi, \omega)v, & \omega(t) &\in R^q, \\ \mu &= c(\xi, \omega) + d(\xi, \omega)v, & v(t) &\in R, \end{aligned}$$

where q is called the dimension of the dynamic state feedback; $a(\xi, \omega)$, $b(\xi, \omega)$ are q -dimensional vector fields; and $c(\xi, \omega)$, $d(\xi, \omega)$ are scalar functions. In general, they are nonlinear.

Consider system (2.4) with a dynamic state feedback (2.14). The extended system is as follows:

$$(2.15) \quad \begin{aligned} \begin{bmatrix} \dot{\xi} \\ \dot{\omega} \end{bmatrix} &= \begin{bmatrix} f(\xi) + g(\xi)c(\xi, \omega) \\ a(\xi, \omega) \end{bmatrix} + \begin{bmatrix} g(\xi)d(\xi, \omega) \\ b(\xi, \omega) \end{bmatrix} v \\ &= f_e(\xi, \omega) + g_e(\xi, \omega)v. \end{aligned}$$

Let F_e be the Jacobian matrix of $f_e(\xi, \omega)$ at $(0, 0)$; let G_e be $g_e(0, 0)$.

DEFINITION 4. If we can find a dynamic state feedback such that the extended system (2.15) is linearly controllable and it can be transformed into

$$(2.16) \quad \dot{z} = F_e z + G_e v + O(z, v)^3$$

by a change of coordinates (in the extended state space)

$$(2.17) \quad \begin{bmatrix} \xi \\ \omega \end{bmatrix} = z + \psi^{[2]}(z),$$

then system (2.4) is called quadratically linearizable by a dynamic state feedback.

THEOREM 4. Any linearly controllable system (2.8) is quadratically linearizable by a dynamic state feedback.

COROLLARY 1. Any linearly controllable system (2.4) is quadratically linearizable by a dynamic state feedback.

In Corollary 2, below, we show that finding a suitable dynamic state feedback and a change of coordinates in the extended space is equivalent to solving a set of linear equations. Suppose that the Taylor series of the vector fields $f(\xi)$ and $g(\xi)$ in system (2.4) are

$$(2.18) \quad f(\xi) = F\xi + f^{[2]}(\xi) + O(\xi)^3, \quad g(\xi) = G + g^{[1]}(\xi) + O(\xi)^2.$$

COROLLARY 2. Suppose that the dimension of the state space of system (2.4) is n . To quadratically linearize this system by a dynamic state feedback, we can use the following $(n-1)$ -dimensional dynamic state feedback:

$$(2.19) \quad \dot{\omega} = A\omega + Bv, \quad \mu = \omega_1 + \gamma^{[1]}(\xi, \omega) + \gamma^{[2]}(\xi, \omega),$$

where (A, B) is in Brunovsky form (2.6a) of dimension $n-1$. The change of coordinates (2.17) in the extended state space is

$$(2.20) \quad \begin{bmatrix} \xi \\ \omega \end{bmatrix} = \begin{bmatrix} z \\ \omega \end{bmatrix} + \begin{bmatrix} \phi^{[2]}(z, \omega_1, \dots, \omega_{n-2}) \\ 0 \end{bmatrix}.$$

The homogeneous polynomials $\gamma^{[1]}(\xi, \omega)$, $\gamma^{[2]}(\xi, \omega)$ and the vector fields $\phi^{[2]}(z, \omega_1, \dots, \omega_{n-2})$ are chosen such that the extended system is linearly controllable and that

$$(2.21) \quad \begin{aligned} & [Fz + G(\omega_1 + \gamma^{[1]}(z, \omega)), \phi^{[2]}(z, \omega_1, \dots, \omega_{n-2})] + \frac{\partial \phi^{[2]}}{\partial \omega} A\omega \\ & = G\gamma^{[2]}(z, \omega) + f^{[2]}(z) + g^{[1]}(z)(\omega_1 + \gamma^{[1]}(z, \omega)). \end{aligned}$$

Furthermore, by (2.19) and (2.20), system (2.4) will be transformed into

$$(2.22) \quad \begin{bmatrix} \dot{z} \\ \dot{\omega} \end{bmatrix} = \begin{bmatrix} Fz + G(\omega_1 + \gamma^{[1]}(z, \omega)) \\ A\omega \end{bmatrix} + \begin{bmatrix} 0 \\ B \end{bmatrix} v + O(z, \omega, v)^3.$$

Remark 2. In Charlet, Lévine, and Marino [4], it was proved that if a single-input system is not exactly linearizable by state feedback, then this system is not linearizable by a dynamic state feedback. The result of Corollary 1 means that in the problem of finding the quadratic linearization, the opposite result is true; i.e., any single-input linearly controllable system is quadratically linearizable by a dynamic state feedback.

The theorems and the corollaries in this section will be proved in § 5.

3. Quadratic equivalence. In this section, we will define the family of all the systems, such as (2.8), of certain dimension to be a linear space. An equivalence relation on this linear space will be introduced. Then several theorems on this equivalence relation and the associated classification will be given. All these results will be used in the proofs of the theorems in § 2, given in § 5. The definition of an equivalent relation can be found in [7].

DEFINITION 5. Consider two systems

$$(3.1a) \quad \dot{\xi} = A\xi + B\mu + f_1^{[2]}(\xi) + g_1^{[1]}(\xi)\mu + O(\xi, \mu)^3,$$

$$(3.1b) \quad \dot{x} = Ax + Bv + f_2^{[2]}(x) + g_2^{[1]}(x)v + O(x, v)^3.$$

System (3.1a) is said to be quadratically state feedback equivalent to system (3.1b) if and only if there exists a change of coordinates and state feedback (2.9) such that system (3.1a) is transformed into

$$(3.2) \quad \dot{x} = Ax + Bv + f_2^{[2]}(x) + g_2^{[1]}(x)v + O(x, v)^3;$$

i.e., system (3.1a) is transformed into a system that agrees with (3.1b) up to an error of third degree.

The first k th terms in the Taylor expansion of a vector field is called a k -jet. Therefore the linear and quadratic parts of system (2.8) is the second jet of this system. Similarly, transformation (2.9) is the second jet of the analytic transformation

$$(3.3) \quad \begin{aligned} \xi &= \xi(x) = x + \phi^{[2]}(x) + O(x)^3, \\ v &= \alpha(x) + \beta(x)\mu = \mu + \alpha^{[2]}(x) + \beta^{[1]}(x)\mu + O(x, \mu)^3. \end{aligned}$$

The family of all the transformations of the form (3.3) is a group. The quotient of this group over the normal subgroup of the transformations with vanishing second jets is also a group, and there is a natural one-to-one correspondence between this quotient group and the family of all the second jet transformations (2.9). Therefore the family of the second jet transformations is also a group. It is denoted by \mathbf{G} . Let T_1 and T_2 be two elements in \mathbf{G} , as follows:

$$(3.4a) \quad T_1: \begin{cases} \xi = \xi_1 + \phi_1^{[2]}(\xi_1), \\ \mu_1 = \mu + \alpha_1^{[2]}(\xi_1) + \beta_1^{[1]}(\xi_1)\mu \end{cases}$$

and

$$(3.4b) \quad T_2: \begin{cases} \xi_1 = \xi_2 + \phi_2^{[2]}(\xi_2), \\ \mu_2 = \mu_1 + \alpha_2^{[2]}(\xi_2) + \beta_2^{[1]}(\xi_2)\mu_1. \end{cases}$$

Then $T_2 \circ T_1$ are the linear and quadratic parts of the composition of the following two transformations:

$$(3.5) \quad T_2 \circ T_1: \begin{cases} \xi = \xi_2 + \phi_1^{[2]}(\xi_2) + \phi_2^{[2]}(\xi_2), \\ \mu_2 = \mu + \alpha_1^{[2]}(\xi_2) + \beta_1^{[1]}(\xi_2)\mu + \alpha_2^{[2]}(\xi_2) + \beta_2^{[1]}(\xi_2)\mu. \end{cases}$$

The inverse of T_1 is

$$(3.6) \quad T_1^{-1}: \begin{cases} \xi_1 = \xi - \phi^{[2]}(\xi), \\ \mu = \mu_1 - \alpha^{[2]}(\xi) - \beta^{[1]}(\xi)\mu_1. \end{cases}$$

That systems (3.1a) and (3.1b) are quadratically state feedback equivalent means that there is an element in the group of second jet transformations \mathbf{G} such that it transforms the second jet of (3.1a) to that of (3.1b). So it is easy to show that quadratic

equivalence is an equivalence relation (see [6]). We can define a classification on the family of all the systems of the form (2.8) by this equivalence relation. Each class of this classification contains all systems that are quadratically state feedback equivalent to each other. In § 5 we prove Theorem 2 by showing that the extended quadratic controller forms are the representatives of all the equivalent classes.

THEOREM 5. *Consider two nonlinear systems*

$$(3.7a) \quad \dot{\xi}_1 = A\xi_1 + B\mu_1 + f_1^{[2]}(\xi_1) + g_1^{[1]}(\xi_1)\mu_1 + O(\xi_1, \mu_1)^3,$$

$$(3.7b) \quad \dot{\xi}_2 = A\xi_2 + B\mu_2 + f_2^{[2]}(\xi_2) + g_2^{[1]}(\xi_2)\mu_2 + O(\xi_2, \mu_2)^3.$$

They are quadratically state feedback equivalent to each other if and only if there exist functions $\alpha^{[2]}(\xi_2)$, $\beta^{[1]}(\xi_2)$, and a vector field $\phi^{[2]}(\xi_2)$ such that

$$(3.8a) \quad [A\xi_2, \phi^{[2]}(\xi_2)] + B\alpha^{[2]}(\xi_2) = f_1^{[2]}(\xi_2) - f_2^{[2]}(\xi_2),$$

$$(3.8b) \quad [B, \phi^{[2]}(\xi_2)] + B\beta^{[1]}(\xi_2) = g_1^{[1]}(\xi_2) - g_2^{[1]}(\xi_2).$$

Proof. These two systems are equivalent if and only if there exists a change of coordinates and state feedback, as follows:

$$(3.9) \quad \xi_1 = \xi_2 + \phi^{[2]}(\xi_2), \quad \mu_1 = \mu_2 + \alpha^{[2]}(\xi_2) + \beta^{[1]}(\xi_2)\mu_2$$

such that (3.7a) is transformed into (3.7b) by (3.9). Substituting (3.9) into (3.7a), we have that

$$(3.10) \quad \begin{aligned} \dot{\xi}_2 = & A\xi_2 + B\mu_2 + f_2^{[2]}(\xi_2) + g_2^{[1]}(\xi_2)\mu_2 + B(\alpha^{[2]}(\xi_2) + \beta^{[1]}(\xi_2)\mu_2) \\ & + f_1^{[2]}(\xi_2) - f_2^{[2]}(\xi_2) - [A\xi_2, \phi^{[2]}(\xi_2)] + g_1^{[1]}(\xi_2)\mu_2 - g_2^{[1]}(\xi_2)\mu_2 \\ & - [B, \phi^{[2]}(\xi_2)]\mu_2 + O(\xi_2, \mu_2)^3. \end{aligned}$$

The detailed proof of (3.10) can be found in Krener et al. [13]. It is clear that (3.10) agrees with (3.7b) up to an error of third and higher degree if and only if equations (3.8) hold. \square

Since the set of all the homogeneous polynomials of (x_1, x_2, \dots, x_n) is a linear space of finite dimension, we can consider $(\phi^{[2]}(x), \alpha^{[2]}(x), \beta^{[1]}(x))$ of (2.9) as an element of a linear space W and $(f^{[2]}(\xi), g^{[1]}(\xi))$ of (2.8) as an element of a linear space V . In this way, we can consider the family of transformation (2.9) and the family of nonlinear system (2.8) as linear spaces W and V . Since the linear part of (2.8) is always in Brunovsky form, we sometimes use $(f^{[2]}, g^{[1]})$ to represent system (2.8). Define a linear map \mathfrak{A} from W to V by the following Lie bracket:

$$(3.11) \quad \mathfrak{A}(\phi^{[2]}(\xi), \alpha^{[2]}(\xi), \beta^{[1]}(\xi)) = ([A\xi, \phi^{[2]}] + B\alpha^{[2]}, [B, \phi^{[2]}] + B\beta^{[1]}).$$

Denote $V_0 = \mathfrak{A}(W)$ = the image of W under \mathfrak{A} . By using these notations, we can rewrite Theorem 5 as follows.

THEOREM 5'. *System (3.7a) is quadratically state feedback equivalent to system (3.7b) if and only if*

$$(3.12) \quad (f_1^{[2]}, g_1^{[1]}) \in (f_2^{[2]}, g_2^{[1]}) + V_0;$$

i.e., $(f_1^{[2]}, g_1^{[1]})$ and $(f_2^{[2]}, g_2^{[1]})$ represent the same element in the quotient space V/V_0 .

Remark 3. Theorem 5' means that there is a one-to-one correspondence between V/V_0 and the family of all equivalent classes.

Remark 4. A special case of Theorem 5' is that system (2.8) is quadratically state feedback equivalent to a linear system if and only if $(f^{[2]}, g^{[1]}) \in V_0$. Therefore the elements of V_0 represent all the systems of the form (2.8) that are quadratically linearizable by (2.9).

The following theorem gives us a geometric necessary and sufficient condition for a system to be quadratically linearizable by the change of coordinates and state feedback (2.9).

THEOREM 6. *Consider system (2.8) and let*

$$(3.13a) \quad X_r = \text{ad}_{A\xi+f^{[2]}(\xi)}^r (B + g^{[1]}(\xi)), \quad 0 \leq r \leq n,$$

$$(3.13b) \quad D^k = C^\infty \text{Span} \{X_r, 0 \leq r < k\}.$$

System (2.8) is quadratically state feedback equivalent to the linear system

$$(3.14) \quad \dot{\xi} = A\xi + B\mu$$

if and only if D^k is first-degree involutive for $k = 1, 2, \dots, n-1$; i.e., for any X and Y in D^k , we have that

$$(3.15) \quad [X, Y] = \sum_{r=0}^{k-1} c_r X_r + O(\xi)^1.$$

Proof. This theorem is a particular case of the theorem in Krener [11].

4. Characteristic numbers. In § 3 we defined an equivalence relation by the change of coordinates and state feedback (2.9). In this section, we answer the question of how to determine whether two systems are quadratically state feedback equivalent without trying to solve the system of equations (3.8). We find a set of numbers associated to system (2.8), called characteristic numbers, so that these numbers are invariant under transformation (2.9). Two systems are quadratically state feedback equivalent if and only if they have the same characteristic numbers.

Let C and H be row vectors such that

$$(4.1a) \quad C = [1, 0, 0, \dots, 0],$$

$$(4.1b) \quad HF^{t-1}G = \begin{cases} 0 & 1 \leq t \leq n-1, \\ 1 & t = n. \end{cases}$$

DEFINITION 6. The characteristic numbers of system (2.4) are

$$(4.2a) \quad a^{tr} = HF^{t-1}[\text{ad}_{f(\xi)}^{r-1}(g(\xi)), \text{ad}_{f(\xi)}^{r-2}(g(\xi))]|_{\xi=0},$$

where

$$(4.2b) \quad 2 \leq r \leq n-1, \quad 1 \leq t \leq n-r.$$

Particularly, the characteristic numbers of system (2.8) are

$$(4.2c) \quad \begin{aligned} a^{tr} &= CA^{t-1}[\text{ad}_{A\xi+f^{[2]}(\xi)}^{r-1}(B + g^{[1]}(\xi)), \text{ad}_{A\xi+f^{[2]}(\xi)}^{r-2}(B + g^{[1]}(\xi))]|_{\xi=0} \\ &= CA^{t-1}[X_{r-1}, X_{r-2}]|_{\xi=0}. \end{aligned}$$

In this section, all the results hold for linearly controllable systems, although they are proved only for the systems whose linear parts are in Brunovsky form.

LEMMA 1. (i) *Let $X(\xi)$ and $Y(\xi)$ be vector fields; then*

$$(4.3) \quad CA^{t-1}[X(\xi), Y(\xi)] = L_X(CA^{t-1}Y) - L_Y(CA^{t-1}X).$$

(ii) *For any integer $r \geq 2$, we have that*

$$(4.4) \quad \begin{aligned} \text{ad}_{A\xi+f^{[2]}(\xi)}^{r-1}(B + g^{[1]}(\xi)) &= (-1)^{r-1}A^{r-1}B + \text{ad}_{A\xi}^{r-1}(g^{[1]}(\xi)) \\ &\quad + \sum_{k=0}^{r-2} \text{ad}_{A\xi}^{r-k-2}[f^{[2]}(\xi), (-1)^k A^k B] + O(\xi)^2. \end{aligned}$$

Proof. (i) It holds that

$$\begin{aligned} CA^{t-1}[X(\xi), Y(\xi)] &= CA^{t-1}\left(\frac{\partial Y}{\partial \xi}X - \frac{\partial X}{\partial \xi}Y\right) \\ &= \frac{\partial CA^{t-1}Y}{\partial \xi}X - \frac{\partial CA^{t-1}X}{\partial \xi}Y \\ &= L_x(CA^{t-1}Y) - L_y(CA^{t-1}X). \end{aligned}$$

(ii) Consider identity (4.4). If $r=2$, then

$$(4.5) \quad ad_{A\xi+f^{[2]}(\xi)}(B+g^{[1]}(\xi)) = -AB + ad_{A\xi}(g^{[1]}(\xi)) + [f^{[2]}(\xi), B] + O(\xi)^2.$$

Therefore identity (4.4) is true for $r=2$. Suppose that (4.4) is correct for $r-1$. Consider that

$$\begin{aligned} &ad_{A\xi+f^{[2]}(\xi)}^{r-1}(B+g^{[1]}(\xi)) \\ &= ad_{A\xi+f^{[2]}(\xi)}\left((-1)^{r-2}A^{r-2}B + ad_{A\xi}^{r-2}(g^{[1]}(\xi))\right. \\ &\quad \left.+ \sum_{k=0}^{r-3} ad_{A\xi}^{r-k-3}[f^{[2]}(\xi), (-1)^k A^k]\right) + O(\xi)^2 \\ &= ad_{A\xi}\left((-1)^{r-2}A^{r-2}B + ad_{A\xi}^{r-2}(g^{[1]}(\xi)) + \sum_{k=0}^{r-3} ad_{A\xi}^{r-k-3}[f^{[2]}(\xi), (-1)^k A^k]\right) \\ (4.6) \quad &+ ad_{f^{[2]}(\xi)}((-1)^{r-2}A^{r-2}B) + O(\xi)^2 \\ &= (-1)^{r-1}A^{r-1}B + ad_{A\xi}^{r-1}(g^{[1]}(\xi)) + \sum_{k=0}^{r-3} ad_{A\xi}^{r-k-2}[f^{[2]}(\xi), (-1)^k A^k B] \\ &\quad + [f^{[2]}(\xi), (-1)^{r-2}A^{r-2}B] + O(\xi)^2 \\ &= (-1)^{r-1}A^{r-1}B + ad_{A\xi}^{r-1}(g^{[1]}(\xi)) \\ &\quad + \sum_{k=0}^{r-2} ad_{A\xi}^{r-k-2}[f^{[2]}(\xi), (-1)^k A^k B] + O(\xi)^2. \end{aligned}$$

Therefore identity (4.4) is true for any $r \geq 2$.

LEMMA 2. The characteristic number a^{tr} is a linear map from V to R ; i.e., a^{tr} is a linear function of $f^{[2]}(\xi)$ and $g^{[1]}(\xi)$.

Proof. By (4.3) and (4.4), we can prove the following identity:

$$\begin{aligned} a^{tr} &= L_{(-1)^{r-1}A^{r-1}B}\left(\sum_{k=0}^{r-3} CA^{t-1}ad_{A\xi}^{r-k-3}[f^{[2]}(\xi), (-1)^k A^k B] + CA^{t-1}ad_{A\xi}^{r-2}(g^{[1]}(\xi))\right) \\ &\quad - L_{(-1)^{r-2}A^{r-2}B}\left(\sum_{k=0}^{r-2} CA^{t-1}ad_{A\xi}^{r-k-1}[f^{[2]}(\xi), (-1)^k A^k B] + CA^{t-1}ad_{A\xi}^{r-1}(g^{[1]}(\xi))\right). \end{aligned}$$

This implies that a^{tr} is a linear function of $f^{[2]}(\xi)$ and $g^{[1]}(\xi)$.

LEMMA 3. A system of the form (2.8) is quadratically linearizable by state feedback if and only if all the characteristic numbers are zero.

Proof. Suppose that a system of the form (2.8) is quadratically linearizable by state feedback. From (4.4) we know that the constant part of the vector fields in D^r is linearly generated by $\{B, AB, A^2B, \dots, A^{r-1}B\}$. From Theorem 6, we know that D^k is first-degree involutive for $k = 1, 2, \dots, n-1$. Therefore

$$(4.7) \quad [X_{r-1}, X_{r-2}] = \sum_{i=1}^r c_i A^{i-1} B + O(\xi)^1.$$

So

$$(4.8) \quad a^{tr} = CA^{t-1}[X_{r-1}, X_{r-2}]|_{\xi=0} = 0, \quad 2 \leq r \leq n-1, \quad 1 \leq t \leq n-r$$

because

$$(4.9) \quad CA^{t-1}A^{k-1}B = 0, \quad 1 \leq k \leq r, \quad 1 \leq t \leq n-r.$$

On the other hand, suppose that

$$(4.10) \quad a^{tr} = 0, \quad 2 \leq r \leq n-1, \quad 1 \leq t \leq n-r;$$

i.e.,

$$(4.11) \quad CA^{t-1}[X_{r-1}, X_{r-2}] = 0, \quad 2 \leq r \leq n-1, \quad 1 \leq t \leq n-r.$$

So

$$(4.12) \quad [X_{r-1}, X_{r-2}] = \sum_{i=1}^r c_i A^{i-1} B + O(\xi)^1$$

for some constants c_i . If D^r is not first-degree involutive, and if D^s is first-degree involutive for any $s < r \leq n-1$, then there exists X_t , $t < r-1$ such that

$$(4.13) \quad [X_{r-1}, X_t] \neq \sum_{i=1}^r d_i A^{i-1} B + O(\xi)^1$$

for any real numbers d_1, d_2, \dots, d_r . By (4.12), we know that

$$(4.14) \quad t < r-2.$$

From the Jacobi identity of Lie bracket, we have that

$$(4.15) \quad [X_{r-1}, X_t] = ad_{A\xi+f^{(2)}(\xi)}([X_{r-2}, X_t]) - [X_{r-2}, X_{t+1}].$$

Since D^{r-1} is first-degree involutive and $t+1 \leq r-2$, we know that

$$(4.16a) \quad [X_{t+1}, X_{r-2}] = \sum_{i=1}^{r-1} \bar{c}_i A^{i-1} B + O(\xi)^1,$$

$$(4.16b) \quad [X_{r-2}, X_t] = \sum_{i=1}^{r-1} \tilde{c}_i A^{i-1} B + O(\xi)^1.$$

This implies that

$$(4.17) \quad [X_{r-1}, X_t] = \sum_{i=1}^r c_i A^{i-1} B + O(\xi)^1.$$

It is a contradiction. So the distribution D^k is first-degree involutive for any $1 \leq k \leq n-1$. This means that the system is quadratically linearizable by state feedback. \square

THEOREM 7. *Two systems of the form (2.8) are quadratically state feedback equivalent if and only if the corresponding characteristics numbers are equal.*

Proof. Consider two systems

$$(4.18a) \quad \dot{\xi}_1 = A\xi_1 + B\mu_1 + f_1^{[2]}(\xi_1) + g_1^{[1]}(\xi_1)\mu_1 + O(\xi_1, \mu_1)^3,$$

$$(4.18b) \quad \dot{\xi}_2 = A\xi_2 + B\mu_2 + f_2^{[2]}(\xi_2) + g_2^{[1]}(\xi_2)\mu_2 + O(\xi_2, \mu_2)^3.$$

Let a_1^{tr} and a_2^{tr} be the characteristic numbers of (4.18a) and (4.18b), respectively.

Suppose that (4.18a) and (4.18b) are quadratically state feedback equivalent. From Theorem 5', we know that

$$(4.19) \quad (f_1^{[2]}, g_1^{[1]}) \in (f_2^{[2]}, g_2^{[1]}) + V_0;$$

i.e.,

$$(4.20a) \quad (f_1^{[2]}, g_1^{[1]}) = (f_2^{[2]}, g_2^{[1]}) + (f^{[2]}, g^{[1]})$$

and

$$(4.20b) \quad (f^{[2]}, g^{[1]}) \in V_0.$$

Let a^{tr} be the characteristic numbers of $(f^{[2]}, g^{[1]})$. Since the characteristic numbers are linear functions of $f^{[2]}$ and $g^{[1]}$ (Lemma 2), we have that

$$(4.21) \quad a_1^{tr} = a_2^{tr} + a^{tr}.$$

From Lemma 3 and (4.20b), we know that $a^{tr} = 0$. So

$$(4.22) \quad a_1^{tr} = a_2^{tr}, \quad 2 \leq r \leq n, \quad 1 \leq t \leq n-r.$$

On the other hand, suppose that all the corresponding characteristic numbers are the same. Then

$$(4.23) \quad a_1^{tr} - a_2^{tr} = 0, \quad 2 \leq r \leq n, \quad 1 \leq t \leq n-r.$$

So

$$(4.24) \quad (f_2^{[2]} - f_1^{[2]}, g_2^{[1]} - g_1^{[1]}) \in V_0.$$

Theorem 5' and (4.24) imply that systems (4.18a) and (4.18b) are quadratically state feedback equivalent.

5. The proofs of the theorems in § 2.

Proof of Theorem 2. Consider the following special kind of $\tilde{f}^{[2]}(x)$:

$$(5.1a) \quad \tilde{f}^{[2]}(x) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \tilde{f}_i(x) \\ \vdots \\ 0 \end{bmatrix} \quad \text{for some } 1 \leq i \leq n;$$

here

$$(5.1b) \quad \tilde{f}_i(x) = a_{ij}x_j^2 \quad \text{for some } j \geq i+2.$$

Then

$$(5.2a) \quad ad_{\tilde{f}^{[2]}(x)}(A^{r-1}B) = 0 \quad \text{for } r \neq n-j+1;$$

$$(5.2b) \quad ad_{\tilde{f}^{[2]}(x)}(A^{n-j}B) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ -2a_{ij}x_j \\ \vdots \\ 0 \end{bmatrix}.$$

Therefore

$$(5.3) \quad ad_{A^{r-1}\tilde{f}^{[2]}(x)}(B) = \begin{cases} (-1)^{r-1}A^{r-1}B & r \leq n-j+1, \\ (-1)^{n-j+1}A^{n-j+1}B + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ (-1)^r 2a_{ij}x_j \\ \vdots \\ 0 \end{bmatrix} + O(x)^2 & r = n-j+2, \\ (-1)^{r-1}A^{r-1}B + \begin{bmatrix} * \\ \vdots \\ * \\ 2a_{ij}x_{2j+r-n-2} \\ \vdots \\ 0 \end{bmatrix} + O(x)^2 & n-j+2 < r \leq 2(n-j+1), \\ (-1)^{r-1}A^{r-1}B + O(x)^2 & r \geq 2(n-j+1). \end{cases}$$

In (5.3), * denotes a linear polynomial of $(x_j, x_{j+1}, \dots, x_n)$. So

$$(5.4a) \quad [ad_{A^{r-1}\tilde{f}^{[2]}(\xi)}(B), ad_{A^{r-2}\tilde{f}^{[2]}(\xi)}(B)] = 0, \quad r \neq n-j+2;$$

$$(5.4b) \quad [ad_{A^{r-1}\tilde{f}^{[2]}(\xi)}(B), ad_{A^{r-2}\tilde{f}^{[2]}(\xi)}(B)] = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 2a_{ij} \\ \vdots \\ 0 \end{bmatrix}, \quad r = n-j+2.$$

Therefore

$$(5.5) \quad \begin{aligned} a^{tr} &= CA^{t-1}[ad_{A^{r-1}\tilde{f}^{[2]}(\xi)}(B), ad_{A^{r-2}\tilde{f}^{[2]}(\xi)}(B)] \\ &= \begin{cases} 2a_{ij} & r = n-j+2 \text{ and } t = i, \\ 0 & \text{others.} \end{cases} \end{aligned}$$

As we know, any $\tilde{f}^{[2]}(x)$ of the form (2.10) is a linear combination of the vector fields given in (5.1). Given any system in the extended quadratic controller form (2.10), from

(5.5) and Lemma 2, we can find the characteristic numbers

$$(5.6) \quad a^{ir} = 2a_{in-r+2}.$$

This implies that, given a set of characteristic numbers, there exists one and only one system in the extended quadratic controller form that has the given characteristic numbers. Theorem 2 follows this fact and Theorem 7.

Proof of Theorem 3. We prove this theorem with the following two steps:

(i) Any two systems given by (2.11) are not quadratically state feedback equivalent to each other.

(ii) Any system is quadratically state feedback equivalent to a system of the form of (2.11).

To prove (i), let us consider the following two systems of the form (2.11):

$$(5.7) \quad (0, \tilde{g}^{[1]}(x)) \quad \text{and} \quad (0, \tilde{g}^{[1]}(x)).$$

They are quadratically state feedback equivalent to each other if and only if

$$(5.8) \quad (0, \tilde{g}^{[1]}(x) - \tilde{g}^{[1]}(x))$$

is quadratically linearizable by state feedback (Theorem 5'). However, (5.8) is also a system of the form (2.1). So proving that the result in part (i) is equivalent to proving that any system (2.11) is not quadratically linearizable by state feedback if $\tilde{g}^{[1]}(x)$ is not zero. Assume that

$$(5.1) \quad \tilde{g}^{[1]}(x) \neq 0$$

and that $\tilde{g}_{t_0}^{[1]}(x)$ is the first entry of $\tilde{g}^{[1]}(x)$ such that $\tilde{g}_t^{[1]}(x) \neq 0$; i.e.,

$$(5.10) \quad \begin{aligned} \tilde{g}_t^{[1]}(x) &= 0 \quad \text{if } t < t_0, \\ \tilde{g}_{t_0}^{[1]}(x) &\neq 0. \end{aligned}$$

Assume that

$$(5.11a) \quad \tilde{g}_{t_0}^{[1]}(x) = a_{n-r+2}x_{n-r+2} + a_{n-r+3}x_{n-r+3} + \cdots + a_nx_n,$$

where

$$(5.11b) \quad a_{n-r+2} \neq 0 \quad \text{and} \quad 2 \leq r \leq t_0.$$

Then we have that

$$(5.12a) \quad \begin{aligned} X_{r-2} &= ad_{Ax}^{r-2}(B + \tilde{g}^{[1]}) \\ &= (-1)^{r-2}A^{r-2}B + (-1)^{r-2} \left\{ \begin{array}{c} 0 \\ \vdots \\ 0 \\ \tilde{g}_{t_0}^{[1]}(x) \\ * \\ \vdots \\ * \end{array} \right\} \left. \vphantom{\begin{array}{c} 0 \\ \vdots \\ 0 \\ \tilde{g}_{t_0}^{[1]}(x) \\ * \\ \vdots \\ * \end{array}} \right\} t_0 - r + 2; \end{aligned}$$

$$(5.12b) \quad \begin{aligned} X_{r-1} &= ad_{Ax}^{r-1}(B + \tilde{g}^{[1]}) \\ &= (-1)^{r-1}A^{r-1}B + (-1)^{r-1} \left\{ \begin{array}{c} 0 \\ \vdots \\ 0 \\ \tilde{g}_{t_0}^{[1]}(x) \\ * \\ \vdots \\ * \end{array} \right\} \left. \vphantom{\begin{array}{c} 0 \\ \vdots \\ 0 \\ \tilde{g}_{t_0}^{[1]}(x) \\ * \\ \vdots \\ * \end{array}} \right\} t_0 - r + 1. \end{aligned}$$

So

$$(5.13) \quad CA^{t_0-r}[X_{r-1}, X_{r-2}] = a_{n-r+2} \neq 0;$$

i.e., the characteristic number a^{t_0-r} is not zero. Therefore $(0, \tilde{g}^{[1]}(x))$ is not quadratically linearizable by state feedback. Part (i) is proved.

Now we can prove part (ii). Since any system is quadratically state feedback equivalent to system (2.10), we must prove that any system (2.10) is quadratically state feedback equivalent to system (2.11). Since any system (2.11) is quadratically state feedback equivalent to exactly one system of the form (2.10) (Theorem 2), and different systems given by (2.11) are quadratically state feedback equivalent to different systems of the form (2.10) (part (i) and Theorem 2), also since the set of systems (2.10) and the set of systems (2.11) are linear space of the same dimension $((n-1)(n-2))/2$, we know that any system (2.10) is quadratically state feedback equivalent to system (2.11). Part (ii) is proved.

Proof of Theorem 4. According to Theorem 2, it is sufficient to prove the result for the systems in the extended quadratic controller form. Let the dynamic state feedback be

$$(5.14) \quad \begin{aligned} \dot{\omega}_1 &= \omega_2, \\ \dot{\omega}_2 &= \omega_3, \\ &\vdots \\ \dot{\omega}_{n-1} &= \tilde{v}, \\ v &= \omega_1 + \gamma^{[2]}(x, \omega), \end{aligned}$$

where $\gamma^{[2]}$ is a homogeneous polynomial of second degree in (x, ω) . The extended system is

$$(5.15) \quad \begin{bmatrix} \dot{x} \\ \dot{\omega} \end{bmatrix} = A_1 \begin{bmatrix} x \\ \omega \end{bmatrix} + B_1 \tilde{v} + \begin{bmatrix} \tilde{f}^{[2]}(x) \\ 0 \end{bmatrix} + \begin{bmatrix} B\gamma^{[2]}(x, \omega) \\ 0 \end{bmatrix} \tilde{v}.$$

Here (A_1, B_1) is in the form of (2.6a) of dimension $2n-1$. We define the change of coordinates as follows:

$$(5.16a) \quad z_1 = x_1,$$

$$(5.16b) \quad z_k = \text{linear and quadratic part of } \dot{z}_{k-1}, \quad 2 \leq k \leq n,$$

$$(5.16c) \quad z_{n+p} = \omega_p, \quad 1 \leq p \leq n-1.$$

We claim that

$$(5.17) \quad z_k = x_k + \psi_k(x, \omega_1, \dots, \omega_{k-2}), \quad 2 \leq k \leq n,$$

where $\psi_k(x, \omega_1, \dots, \omega_{k-2})$ is a homogeneous polynomial of second degree. For $k=2$, we have that

$$(5.18) \quad z_2 = \text{linear and quadratic part of } \dot{z}_1 = x_2 + \sum_{j \geq 3}^n a_{2j} x_j^2.$$

So (5.17) is true. Assume that (5.17) is true for $k-1$; then

$$(5.19) \quad \begin{aligned} \dot{z}_{k-1} &= \dot{x}_{k-1} + \dot{\psi}_{k-1}(x, \omega_1, \dots, \omega_{k-3}) \\ &= x_k + \sum_{j \geq k+1} a_{k-1j} x_j^2 + \dot{\psi}_{k-1}(x, \omega_1, \dots, \omega_{k-3}) \\ &= x_k + \psi_k(x, \omega_1, \dots, \omega_{k-2}) + O(x, \omega_1, \dots, \omega_{k-2})^3. \end{aligned}$$

The last equality is true because

$$(5.20) \quad \dot{x}_i = x_{i+1} + \tilde{f}^{[2]}(x) + O(x)^3, \quad \dot{x}_n = \omega_1 + \gamma^{[2]}(x, \omega),$$

and $\dot{\omega}_1, \dots, \dot{\omega}_{k-3}$ are related only to $\omega_1, \dots, \omega_{k-2}$. Therefore $z_k = x_k + \psi_k(x, \omega_1, \dots, \omega_{k-2})$. So (5.17) is true for any $2 \leq k \leq n$. By (5.16) and (5.17), we have that

$$(5.21a) \quad \begin{aligned} \dot{z}_1 &= z_2, \\ \dot{z}_2 &= z_3 + O(z, \tilde{v})^3, \\ &\vdots \\ \dot{z}_{n-1} &= z_n + O(z, \tilde{v})^3, \end{aligned}$$

and

$$(5.21b) \quad \begin{aligned} \dot{z}_n &= \dot{x}_n + \dot{\psi}_n(x, \omega_1, \dots, \omega_{n-2}) \\ &= \omega_1 + \gamma^{[2]}(x, \omega) + \dot{\psi}_n(x, \omega_1, \dots, \omega_{n-2}). \end{aligned}$$

Let

$$(5.22) \quad \gamma^{[2]} = \text{the quadratic part of } -\dot{\psi}_n(x, \omega_1, \dots, \omega_{n-2});$$

then

$$(5.23) \quad \dot{z}_n = \omega_1 = z_{n+1} + O(z, \tilde{v})^3.$$

Therefore, by the change of coordinates (5.16) and (5.22), system (5.15) is transformed into

$$(5.24) \quad \begin{aligned} \dot{z}_1 &= z_2, \\ \dot{z}_2 &= z_3 + O(z, \tilde{v})^3, \\ &\vdots \\ \dot{z}_{n-1} &= z_n + O(z, \tilde{v})^3, \\ \dot{z}_n &= z_{n+1} + O(z, \tilde{v})^3, \\ \dot{z}_{n+1} &= z_{n+2}, \\ &\vdots \\ \dot{z}_{2n-1} &= \tilde{v}. \end{aligned}$$

It is linearly controllable system without quadratic terms. Theorem 4 is proved.

Remark 5. Sometimes the dimension of the dynamic state feedback used in Theorem 2 can be less than $n-1$. Suppose that a system is in extended quadratic controller form (2.10). Let

$$(5.25) \quad q = \max \{j-i; a_{ij} \neq 0, j \geq i+2, 1 \leq i \leq n-2\}.$$

To quadratically linearize the system, a q -dimensional dynamic state feedback is sufficient. The proof is almost the same as above, except that (5.17) is changed to

$$(5.26) \quad z_k = x_k + \psi_k(x, \omega_1, \dots, \omega_{k-1-n+q}), \quad n-q+1 \leq k \leq n.$$

Remark 6. From this proof, we find that the dynamic state feedback is chosen to be in Brunovsky form, as follows:

$$(5.27) \quad \dot{\omega} = A\omega + Bv, \quad \mu = \omega_1 + \gamma^{[2]}(x, \omega).$$

Furthermore, (5.16) and (5.17) imply that the change of coordinates in the extended state space is

$$(5.28) \quad \begin{bmatrix} x \\ \omega \end{bmatrix} = \begin{bmatrix} z \\ \omega \end{bmatrix} + \begin{bmatrix} \phi^{[2]}(z, \omega_1, \dots, \omega_{n-2}) \\ 0 \end{bmatrix};$$

i.e., ω is not changed, and the quadratic part is independent of ω_{n-1} .

Proof of Corollary 1. By Theorem 1, there exists a linear change of coordinates and state feedback

$$(5.29) \quad \xi_1 = T\xi, \quad \mu_1 = \alpha(\xi_1) + \beta\mu,$$

where $\alpha(\xi_1)$ is a linear function and $\beta \neq 0$ is a constant, such that system (2.4) is transformed into

$$(5.30) \quad \begin{aligned} \dot{\xi}_1 &= Tf(T^{-1}\xi_1) + Tg(T^{-1}\xi_1) \left(\frac{1}{\beta}\mu_1 - \frac{\alpha(\xi_1)}{\beta} \right) \\ &= A\xi_1 + B\mu_1 + f^{[2]}(\xi_1) + g^{[1]}(\xi_1)\mu_1 + O(\xi_1, \mu_1)^3, \end{aligned}$$

where (A, B) is in Brunovsky form. By Theorem 4, we can find a dynamic state feedback such that the extended system can be linearized to the second degree by a change of coordinates. This extended system is

$$(5.31) \quad \begin{aligned} \dot{\xi}_1 &= Tf(T^{-1}\xi_1) + Tg(T^{-1}\xi_1) \left(\frac{1}{\beta}\mu_1 - \frac{\alpha(\xi_1)}{\beta} \right), \\ \dot{\omega} &= a(\xi_1, \omega) + b(\xi_1, \omega)v, \\ \mu_1 &= c(\xi_1, \omega) + d(\xi_1, \omega)v. \end{aligned}$$

It is linearly controllable. Under the old coordinates ξ , this system is

$$(5.32) \quad \begin{aligned} \dot{\xi} &= f(\xi) + g(\xi) \left\{ \frac{1}{\beta}(c(T^{-1}\xi, \omega) + d(T^{-1}\xi, \omega)v) - \frac{\alpha(T^{-1}\xi)}{\beta} \right\}, \\ \dot{\omega} &= a(T^{-1}\xi, \omega) + b(T^{-1}\xi, \omega)v. \end{aligned}$$

If we define

$$(5.33) \quad \mu = \frac{1}{\beta}(c(T^{-1}\xi, \omega) + d(T^{-1}\xi, \omega)v) - \frac{\alpha(T^{-1}\xi, \omega)}{\beta}$$

as the output of the dynamic state feedback, then (5.32) becomes

$$(5.34) \quad \begin{aligned} \dot{\xi} &= f(\xi) + g(\xi)\mu, \\ \dot{\omega} &= a(T^{-1}\xi, \omega) + b(T^{-1}\xi, \omega)v, \\ \mu &= \frac{1}{\beta}(c(T^{-1}\xi, \omega) + d(T^{-1}\xi, \omega)v) - \frac{\alpha(T^{-1}\xi)}{\beta}. \end{aligned}$$

System (5.34) is quadratically linearizable under a change of coordinates because system (5.32) is quadratically linearizable. This implies that the system

$$(5.35) \quad \dot{\xi} = f(\xi) + g(\xi)\mu$$

is quadratically linearizable by a dynamic state feedback. Corollary 1 is proved.

Proof of Corollary 2. By Remark 6, we know that the dynamic state feedback in (5.31) can be chosen in Brunovsky form, as follows:

$$(5.36) \quad \dot{\omega} = A\omega + Bv, \quad \mu_1 = \omega_1 + \gamma^{[2]}(\xi_1, \omega).$$

The dimension of A and B is $n-1$. In this case, the dynamic state feedback in (5.34) is

$$(5.37) \quad \dot{\omega} = A\omega + Bv, \quad \mu = \frac{1}{\beta}\omega_1 - \frac{\alpha(T^{-1}\xi)}{\beta} + \gamma^{[2]}(T^{-1}\xi, \omega).$$

We make the change of coordinates for ω and v and denote $(1/\beta)\omega$ by ω and $(1/\beta)v$ by v ; then the dynamic state feedback (5.37) will be changed to a dynamic state feedback that is in the same form as (2.19). Therefore we proved that any system (2.4) is quadratically linearizable by the dynamic state feedback (2.19). By using Taylor's series expansion (2.18), the extended system is

$$(5.38) \quad \begin{aligned} \dot{\xi} &= F\xi + G(\omega_1 + \gamma^{[1]}(\xi, \omega) + \gamma^{[2]}(\xi, \omega)) \\ &\quad + f^{[2]}(\xi) + g^{[1]}(\xi)(\omega_1 + \gamma^{[1]}(\xi, \omega)) + O(\xi, \omega)^3, \\ \dot{\omega} &= A\omega + Bv. \end{aligned}$$

This system is linearizable by a change of coordinates. From Remark 6, we know that the change of coordinates can be chosen in the form of (5.28); i.e.,

$$(5.39) \quad \begin{bmatrix} \xi \\ \omega \end{bmatrix} = \begin{bmatrix} z \\ \omega \end{bmatrix} + \begin{bmatrix} \phi^{[2]}(z, \omega_1, \dots, \omega_{n-2}) \\ 0 \end{bmatrix}.$$

Substituting this into the equations in the Theorem 5, we have that

$$(5.40a) \quad \begin{aligned} &\begin{bmatrix} \left(Fz + G(\omega_1 + \gamma^{[1]}(z, \omega)) \right) \\ A\omega \end{bmatrix}, \begin{bmatrix} \phi^{[2]}(z, \omega_1, \dots, \omega_{n-2}) \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} G\gamma^{[2]}(z, \omega) + f^{[2]}(z) + g^{[1]}(z)\omega_1 + g^{[1]}(z)\gamma^{[1]}(z, \omega) \\ 0 \end{bmatrix}, \end{aligned}$$

$$(5.40b) \quad \begin{bmatrix} \begin{pmatrix} 0 \\ B \end{pmatrix}, \begin{bmatrix} \phi^{[2]}(z, \omega_1, \dots, \omega_{n-2}) \\ 0 \end{bmatrix} \end{bmatrix} = 0.$$

Since $(\partial\phi^{[2]}(z, \omega_1, \dots, \omega_{n-2}))/\partial\omega_{n-1} = 0$, (5.40b) is always true. Equation (5.40a) is equivalent to

$$(5.41) \quad \begin{aligned} &\begin{bmatrix} \frac{\partial\phi^{[2]}}{\partial z}(Fz + G\omega_1 + G\gamma^{[1]}) + \frac{\partial\phi^{[2]}}{\partial\omega}A\omega \\ 0 \end{bmatrix} - \begin{bmatrix} F\phi^{[2]} + G\frac{\partial\gamma^{[1]}}{\partial z}\phi^{[2]} \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} G\gamma^{[2]}(z, \omega) + f^{[2]}(z) + g^{[1]}(z)(\omega_1 + \gamma^{[1]}(z, \omega)) \\ 0 \end{bmatrix}. \end{aligned}$$

It is equivalent to

$$(5.42) \quad \begin{aligned} &[Fz + G(\omega_1 + \gamma^{[1]}(z, \omega)), \phi^{[2]}(z, \omega_1, \dots, \omega_{n-2})] + \frac{\partial\phi^{[2]}(z, \omega_1, \dots, \omega_{n-2})}{\partial\omega}A\omega \\ &= G\gamma^{[2]}(z, \omega) + f^{[2]}(z) + g^{[1]}(z)(\omega_1 + \gamma^{[1]}(z, \omega)). \end{aligned}$$

Corollary 2 is proved.

Remark 7. From (5.37) we know that $\gamma^{[1]}(z, \omega)$ can be chosen as $-\alpha(T^{-1}\xi)/\beta$.

6. An example of Theorem 4. Consider that

$$(6.1) \quad \dot{\xi}_1 = \xi_2 + \xi_3^2, \quad \dot{\xi}_2 = \xi_3, \quad \dot{\xi}_3 = \mu.$$

This is a system in extended quadratic controller form, so it is a typical three-dimensional system that is not quadratically linearizable by state feedback. We construct the following dynamic state feedback:

$$(6.2) \quad \dot{\xi}_4 = \xi_5, \quad \dot{\xi}_5 = v, \quad \mu = \xi_4 + \gamma^{[2]}(\xi_1, \xi_2, \xi_3, \xi_4, \xi_5),$$

where $\gamma^{[2]}(\xi_1, \xi_2, \xi_3, \xi_4, \xi_5)$ is a quadratic homogeneous polynomial in $(\xi_1, \xi_2, \xi_3, \xi_4, \xi_5)$, which will be determined later. The extended system is

$$(6.3) \quad \begin{aligned} \dot{\xi}_1 &= \xi_2 + \xi_3^2, \\ \dot{\xi}_2 &= \xi_3, \\ \dot{\xi}_3 &= \xi_4 + \gamma^{[2]}(\xi_1, \xi_2, \xi_3, \xi_4, \xi_5), \\ \dot{\xi}_4 &= \xi_5, \\ \dot{\xi}_5 &= v. \end{aligned}$$

By Hunt and Su's method of linearization, let us take

$$(6.4) \quad \begin{aligned} z_1 &= \xi_1, \\ z_2 &= \xi_2 + \xi_3^2 = \dot{z}_1, \\ z_3 &= \xi_3 + 2\xi_3\xi_4 = \text{linear and quadratic part of } \dot{z}_2, \\ z_4 &= \xi_4 + \gamma + 2\xi_4^2 + 2\xi_3\xi_5 = \text{linear and quadratic part of } \dot{z}_3, \\ z_5 &= \xi_5. \end{aligned}$$

If we take $\gamma^{[2]} = -2\xi_4^2 - 2\xi_3\xi_5$, then

$$(6.5) \quad z_4 = \dot{z}_4.$$

Therefore we have that

$$(6.6) \quad \begin{aligned} \dot{z}_1 &= x_2, \\ \dot{z}_2 &= x_3 + O(x, v)^3, \\ \dot{z}_3 &= x_4 + O(x, v)^3, \\ \dot{z}_4 &= x_5, \\ \dot{z}_5 &= v. \end{aligned}$$

Therefore system (6.1) is quadratically linearizable by the dynamic state feedback (6.2). This is an example of Theorem 4. In fact, the idea used in the proof of Theorem 4 is similar to the argument in this example.

In this paper, all the results are restricted to the single-input nonlinear systems. In fact, similar results in the multi-input case are also correct, and they will be given in another paper. The idea of finding quadratic normal forms and extending the state space was also successfully used in the problem of finding nonlinear observers.

REFERENCES

- [1] R. W. BROCKETT, *Feedback invariants for nonlinear systems*, in Proc. IFAC Congress, Helsinki, 1978.
- [2] P. BRUNOVSKY, *A classification of linear controllable systems*, Kybernetika, 3 (1970), pp. 173–187.
- [3] B. CHARLET, J. LÉVINE, AND R. MARINO, *On dynamic feedback linearization*, Systems Control Lett., 13 (1989), pp. 143–152.
- [4] ———, *Dynamic feedback linearization and applications to aircraft control*, in Proc. IEEE Conf. on Decision and Control, Austin, TX, 1988, pp. 701–705.
- [5] L. R. HUNT AND R. SU, *Linear equivalent of nonlinear time varying systems*, in Proc. MTNS, Santa Monica, CA, 1981, pp. 119–123.
- [6] A. ISIDORI, *Nonlinear Control Systems*, 2nd ed., Springer-Verlag, Berlin, New York, 1989.
- [7] N. JACOBSON, *Basic Algebra*, W. H. Freeman, New York, 1985.
- [8] B. JAKUBCZYK AND W. RESPONDEK, *On the linearization of control systems*, Bull. Acad. Polon. Sci. Ser. Math. Astronom. Phys., 28 (1980), pp. 517–522.
- [9] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
- [10] S. KARAHAN, *Higher order linear approximation to nonlinear systems*, Ph.D. thesis, Dept. of Mechanical Engineering, University of California, Davis, CA, 1988.
- [11] A. J. KRENER, *Approximate linearization by state feedback and coordinate change*, Systems Control Lett., 5 (1984), pp. 181–185.
- [12] ———, *Normal forms for linear and nonlinear systems*, in Differential Geometry, The Interface between Pure and Applied Mathematics, M. Luksik, C. Martin, and W. Shadwick, eds., Contemporary Mathematics, Vol. 68, American Mathematical Society, Providence, RI, 1986, pp. 157–189.
- [13] A. J. KRENER, S. KARAHAN, M. HUBBARD, AND R. FREZZA, *Higher order linear approximations to nonlinear control systems*, in Proc. IEEE Conf. on Decision and Control, Los Angeles, CA, 1987, pp. 519–523.
- [14] A. PHELPS AND A. J. KRENER, *Computation of observer normal forms using MACSYMA*, in Nonlinear Dynamics and Control, C. Byrnes, C. Martin, and R. Sacks, eds., North-Holland, Amsterdam, 1988.
- [15] S. N. SINGH, *Decoupling of invertible nonlinear systems with state feedback and precompensation*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 1237–1239.
- [16] R. SOMMER, *Control design for multivariable nonlinear time-varying systems*, Internat. J. Control, 31 (1980), pp. 883–891.

ON THE EXPONENTIAL STABILITY OF SINGULARLY PERTURBED SYSTEMS*

MARTIN CORLESS† AND LUIGI GLIELMO‡

Abstract. This paper establishes some results and properties related to the exponential stability of general dynamical systems and, in particular, singularly perturbed systems. For singularly perturbed systems it is shown that if both the reduced-order system and the boundary-layer system are exponentially stable, then, provided that some further regularity conditions are satisfied, the full-order system is exponentially stable for sufficiently small values of the perturbation parameter μ , and its rate of convergence approaches that of the reduced-order system ($\mu = 0$) as μ approaches zero. Exponentially decaying norm bounds are given for the “slow” and “fast” components of the full-order system trajectories. To achieve this result, a new converse Lyapunov result for exponentially stable systems is presented.

Key words. singularly perturbed systems, exponential stability, Lyapunov stability, converse Lyapunov results

AMS(MOS) subject classifications. 34D15, 34D20, 34E15, 93D05

1. Introduction. Recently, considerable attention has been devoted to the study of singularly perturbed systems and, in particular, to their stability properties (see [10], [11], and the references therein). Loosely speaking, a common problem considered in the literature is as follows. Given the stability characteristics of the two limiting systems obtained by letting the perturbation parameter equal zero in the “slow” timescale and the “fast” timescale, i.e., the reduced-order system and the boundary-layer system, respectively, determine the stability characteristics of the full-order system when the perturbation parameter is sufficiently small but nonzero. Various approaches to this problem have been suggested in the literature; here, as in [12], we make use of Lyapunov functions [9]. In this framework, two Lyapunov functions are considered, one for the reduced-order system and one for the boundary-layer system. Then, viewing the full-order system as an interconnected system, a candidate for a so-called “composite” Lyapunov function is constructed.

Our attention is focused on the property of exponential stability; we show that if both the reduced-order system and the boundary-layer system are exponentially stable and some further regularity conditions are satisfied, then the full-order system is exponentially stable for sufficiently small values of the perturbation parameter. This result is already present in [12]. The emphasis here is on the estimation of the rate of convergence of the full-order system. We prove the following. Consider any rate of exponential convergence that is less than the supremal rate of convergence of the reduced-order system. Then the full-order system has a rate of convergence that is a continuous function of the singular perturbation parameter (when this parameter is sufficiently small), and this rate approaches the chosen rate of the reduced-order system as the perturbation parameter approaches zero.

Moreover, the norm of the trajectories of the boundary-layer state are shown to be bounded above by the sum of two exponentially decaying terms, one in the fast

* Received by the editors August 6, 1990; accepted for publication (in revised form) May 31, 1991.

† School of Aeronautics and Astronautics, Purdue University, West Lafayette, Indiana 47907. This author was supported by U.S. National Science Foundation grants MSM-87-06927 and MSS-90-57079.

‡ Dipartimento di Informatica e Sistemistica, Università degli Studi di Napoli “Federico II,” via Claudio 21, 80125 Napoli, Italy. This research was performed during a stay at the School of Aeronautics and Astronautics, Purdue University and was supported by Consiglio Nazionale delle Ricerche, Italy grant 203.07.17.

timescale and one in the slow timescale. The rates of convergence of these terms possess properties analogous to those described above.

To obtain these results, we develop a converse Lyapunov result for exponentially stable systems (Lemma 1), which we consider interesting in itself.

Two examples are presented to illustrate these results.

1.1. Notation. The following notation will be employed in the paper:

$\mathbf{R} (\mathbf{R}_+)$	the set of (nonnegative) real numbers;
I_n	the $n \times n$ identity matrix (the subscript will be dropped when n is clear from the context);
$\sigma(A)$	the set of eigenvalues of the square matrix A ;
$\Re(\lambda)$	the real part of the complex number λ ;
$\lambda_{\min}[Q]$	the minimum eigenvalue of the symmetric matrix Q ;
$D_i f$	the "block" partial derivative of the function f with respect to its i th argument [1, p. 360];
$\text{blockdiag} (A_1, \dots, A_n)$	the block-diagonal matrix whose diagonal elements are the matrices A_1, \dots, A_n .

In addition, when talking about general properties of parameterized dynamical systems, we will refer to a system described by the parameterized differential equation

$$(\Sigma) \quad \dot{\xi}(t) = p(t, \xi(t), \theta),$$

where $t \in \mathbf{R}$ is the "time," $\xi(t) \in \mathbf{R}^{n_\xi}$ is the state vector, and θ is a parameter vector ranging in some nonempty set Θ . A solution, corresponding to an initial condition $\xi(t_0) = \xi_0$, will be denoted by $\phi(\cdot; t_0, \xi_0, \theta)$ or, when no confusion is likely to arise, simply by $\xi(\cdot)$.

Given a function $L: (t, \xi, \theta) \mapsto L(t, \xi, \theta)$, $L_{(\Sigma)}(t)$ will mean the value taken at time t by the function L along a trajectory of system (Σ) , i.e., $L_{(\Sigma)}(t) \triangleq L(t, \phi(t; t_0, \xi_0, \theta), \theta)$. The dependence on t_0 , ξ_0 , and θ is omitted for the sake of brevity. Since $\dot{L}_{(\Sigma)}(t) = D_1 L(t, \xi(t), \theta) + D_2 L(t, \xi(t), \theta) p(t, \xi(t), \theta)$, we will sometimes write this as $\dot{L}_{(\Sigma)}(t, \xi, \theta)$.

We will frequently use the fact that, if $f: \mathbf{R}^p \rightarrow \mathbf{R}^q$ is continuously differentiable and $\|Df(x)\| \leq M$ for all $x \in \mathbf{R}^p$, then

$$\|f(y) - f(x)\| \leq M \|y - x\|$$

for all $x, y \in \mathbf{R}^p$ [1, Cor. 40.6].

Arguments of functions will sometimes be omitted if this is not likely to cause confusion.

2. The main result. We consider the singularly perturbed system

$$(1a) \quad \dot{x}(t) = f(t, x(t), z(t), \mu),$$

$$(1b) \quad \mu \dot{z}(t) = g(t, x(t), z(t), \mu),$$

where $t \in \mathbf{R}$ is the "time," $x(t) \in \mathbf{R}^n$ and $z(t) \in \mathbf{R}^m$ are the state variables, and $\mu > 0$ is the singular perturbation parameter. For some $\bar{\mu} > 0$, the functions $f: \mathbf{R} \times \mathbf{R}^n \times \mathbf{R}^m \times [0, \bar{\mu}] \rightarrow \mathbf{R}^n$ and $g: \mathbf{R} \times \mathbf{R}^n \times \mathbf{R}^m \times [0, \bar{\mu}] \rightarrow \mathbf{R}^m$ are continuous. Before stating our assumptions on system (1), we need the following definition.

DEFINITION 1. A parameterized dynamical system $\dot{\xi} = p(t, \xi, \theta)$ is *globally uniformly exponentially stable* (g.u.e.s.) if and only if there exist positive scalars c and α such that, for all $t_0 \in \mathbf{R}$, $\xi_0 \in \mathbf{R}^{n_\xi}$, $\theta \in \Theta$, and $t \geq t_0$,

$$\|\xi(t)\| \leq c \|\xi_0\| e^{-\alpha(t-t_0)},$$

where $\xi(t) = \phi(t; t_0, \xi_0, \theta)$.

The scalars c and α will be called a *gain* and a *rate of exponential convergence*, respectively. Note that in the above definition they are independent of θ . The supremum $\bar{\alpha}$ of all the rates of exponential convergence will be called the *supremal rate of exponential convergence*; this may or may not be an actual convergence rate. An immediate consequence of the above property is that $\xi = 0$ is an equilibrium state for all θ . If for each θ , the system has an equilibrium state ξ_θ that is not necessarily zero, we say that the system is g.u.e.s. (about ξ_θ) if it satisfies the requirements of the above definition with $\xi(t)$ and ξ_0 replaced by $\xi(t) - \xi_\theta$ and $\xi_0 - \xi_\theta$, respectively.

Regarding system (1), we first assume the following.

Assumption 1. For each $t \in \mathbf{R}$ and $x \in \mathbf{R}^n$, the equation $0 = g(t, x, \bar{z}, 0)$ has a unique solution $\bar{z} = h(t, x)$, and h is continuously differentiable.

This assumption allows us to uniquely define the *reduced-order system*, sometimes also called the *degenerate system* [7], by setting $\mu = 0$ in (1), as follows:

$$(2a) \quad \dot{x} = f(t, x, z, 0),$$

$$(2b) \quad 0 = g(t, x, z, 0).$$

We see that the second differential equation from (1) has become an “algebraic equation.” In view of Assumption 1, (2b) has a unique solution $z = h(t, x)$, which, upon substitution into (2a), yields the reduced-order system

$$(3) \quad \dot{x} = \bar{f}(t, x),$$

with

$$(4) \quad \bar{f}(t, x) \triangleq f(t, x, h(t, x), 0).$$

In the following, system (2) will be referred to as the *complete reduced-order system*. We assume the following for the reduced-order system.

Assumption 2. The reduced-order system (3) is g.u.e.s. with supremal rate of exponential convergence $\bar{\alpha}_x$.

It follows from Assumption 2 that $\bar{f}(t, 0) \equiv 0$.

To define the *boundary-layer system*, consider any $t_0 \in \mathbf{R}$ and define the “fast time” variable $\tau \triangleq (t - t_0)/\mu$ and new τ -dependent state variables

$$(5) \quad x_f(\tau) \triangleq x(t_0 + \mu\tau) = x(t),$$

$$(6) \quad z_f(\tau) \triangleq z(t_0 + \mu\tau) = z(t).$$

Equations (1) yield

$$(7a) \quad \frac{dx_f}{d\tau}(\tau) = \mu f(t_0 + \mu\tau, x_f(\tau), z_f(\tau), \mu),$$

$$(7b) \quad \frac{dz_f}{d\tau}(\tau) = g(t_0 + \mu\tau, x_f(\tau), z_f(\tau), \mu).$$

Letting $\mu = 0$ in (7), the first equation becomes $dx_f/d\tau \equiv 0$, which implies that $x_f(\tau) \equiv x_f(0) = x(t_0) \triangleq x_0$. Thus the boundary-layer system is described by

$$(8) \quad \frac{dz_f}{d\tau}(\tau) = g(t_0, x_0, z_f(\tau), 0).$$

Note that t_0 and x_0 are treated as parameters in (8).

We assume the following for the boundary-layer system.

Assumption 3. The boundary-layer system (8) (with (t_0, x_0) as a parameter vector) is g.u.e.s. about $h(t_0, x_0)$ with supremal rate of convergence $\bar{\alpha}_y$.

Assumptions 1–3 are the main ones. We require g.u.e.s. of the two systems obtained by letting $\mu = 0$: the reduced-order system in the “slow” timescale and the boundary-layer system in the “fast” timescale. The remaining two assumptions place some regularity conditions on f , g , h .

For convenience in stating our remaining assumptions and our main results, we replace state z by

$$(9) \quad y \triangleq z - h(t, x);$$

we call y the boundary-layer state. Describing system (1) in terms of the state (x, y) we obtain

$$(10a) \quad \dot{x}(t) = F(t, x(t), y(t), \mu),$$

$$(10b) \quad \mu \dot{y}(t) = G(t, x(t), y(t), \mu),$$

with

$$(11) \quad F(t, x, y, \mu) \triangleq f(t, x, h(t, x) + y, \mu),$$

$$(12) \quad G(t, x, y, \mu) \triangleq g(t, x, h(t, x) + y, \mu) - \mu[D_1 h(t, x) + D_2 h(t, x)F(t, x, y, \mu)].$$

Note that the reduced-order system and the boundary-layer system can now be described by

$$(13) \quad \dot{x}(t) = F(t, x(t), 0, 0),$$

$$(14) \quad \frac{dy_f}{d\tau}(\tau) = G(t_0, x_0, y_f(\tau), 0),$$

respectively.

Assumption 4. The function f is continuously differentiable with respect to x and z , the function g is continuously differentiable, and there exists a real number $M \geq 0$ such that, for all $t \in \mathbf{R}$, $x \in \mathbf{R}^n$ and $z \in \mathbf{R}^m$,

$$(15) \quad \|D_i f(t, x, z, 0)\|, \|D_i g(t, x, z, 0)\|, \|D_2 h(t, x)\| \leq M, \quad i = 2, 3,$$

$$(16) \quad \|D_1 g(t, x, z, 0)\| \leq M(\|x\| + \|y\|),$$

$$(17) \quad \|D_1 h(t, x)\| \leq M\|x\|.$$

Assumption 5. There exist continuous functions $k_f, k_g: [0, \bar{\mu}] \rightarrow \mathbf{R}_+$, with $k_f(0) = k_g(0) = 0$, and a positive constant d_g such that, for all $t \in \mathbf{R}$, $x \in \mathbf{R}^n$, $z \in \mathbf{R}^m$, and $\mu \in (0, \bar{\mu})$,

$$(18) \quad \|f(t, x, z, \mu) - f(t, x, z, 0)\| \leq k_f(\mu)(\|x\| + \|y\|),$$

$$(19) \quad \|g(t, x, z, \mu) - g(t, x, z, 0)\| \leq k_g(\mu)(\|x\| + \|y\|),$$

$$(20) \quad k_g(\mu)/\mu \leq d_g.$$

Remark 1. Assumption 4 guarantees smooth behavior of f , g , h with respect to t , x , y . Assumption 5 guarantees that the dependency of f and g on μ goes to zero as (x, y) goes to zero. Even for regularly perturbed systems, nonsatisfaction of this condition can destroy exponential stability. Consider, for example, the scalar system

$$\dot{x} = -x + \mu(t^{-1} - t^{-2}), \quad t \geq 1,$$

with initial condition $x(1) = x_0$. Its solutions are given by

$$x(t) = e^{-(t-1)}(x_0 - \mu) + \mu t^{-1}.$$

This system is exponentially stable for $\mu = 0$. For $\mu \neq 0$, all solutions converge to 0, but the system is not exponentially stable.

Remark 2. We note that if system (1) is time-invariant, inequalities (16), (17) in Assumption 4 are trivially satisfied; similarly, Assumption 5 is trivially satisfied if f and g do not depend on μ . In particular, for linear time-invariant singularly perturbed systems of the form

$$\dot{x} = A_{11}x + A_{12}z, \quad \mu \dot{z} = A_{21}x + A_{22}z,$$

Assumptions 4 and 5 are always satisfied.

The x and y components of a solution of (10), with initial conditions $x(t_0) = x_0$, $y(t_0) = y_0$, will be denoted by $\phi_x(\cdot; t_0, x_0, y_0)$ and $\phi_y(\cdot; t_0, x_0, y_0)$, respectively. We point out that these solutions also depend on the parameter μ , but we prefer not to explicitly denote this to avoid cumbersome notation. We are now able to state the main results of this paper.

THEOREM 1. *Suppose that Assumptions 1–5 hold and consider any positive $\alpha_x < \bar{\alpha}_x$ and $\alpha_y < \bar{\alpha}_y$. Then there exist positive constants μ^* , c_1 , and continuous functions α_s , $\alpha_f: (0, \mu^*) \rightarrow (0, \infty)$, $\gamma: (0, \mu^*) \rightarrow \mathbf{R}_+$ such that the following hold for all $t_0 \in \mathbf{R}$, $x_0 \in \mathbf{R}^n$, and $y_0 \in \mathbf{R}^m$.*

1. *For each $\mu \in (0, \mu^*)$ and $t \geq t_0$, the trajectories $\phi_x(\cdot; t_0, x_0, y_0)$ and $\phi_y(\cdot; t_0, x_0, y_0)$ of (10) are bounded as follows:*

$$(21) \quad \|\phi_x(t; t_0, x_0, y_0)\| \leq c_1[\|x_0\| + \gamma(\mu)\|y_0\|] e^{-\alpha_s(\mu)(t-t_0)},$$

$$(22) \quad \|\phi_y(t; t_0, x_0, y_0)\| \leq c_1\|y_0\| e^{-\alpha_f(\mu)(t-t_0)/\mu} + \mu c_1[\|x_0\| + \gamma(\mu)\|y_0\|] e^{-\alpha_s(\mu)(t-t_0)};$$

2. *As $\mu \rightarrow 0$,*

$$(23) \quad \alpha_s(\mu) \rightarrow \alpha_x,$$

$$(24) \quad \alpha_f(\mu) \rightarrow \alpha_y,$$

$$(25) \quad \gamma(\mu) \rightarrow 0.$$

The following corollary is a straightforward consequence.

COROLLARY 1. *If Assumptions 1–5 hold, then for any positive $\alpha_x < \bar{\alpha}_x$, there exist a positive constant μ^* and a positive continuous function $\alpha_s: (0, \mu^*) \rightarrow \mathbf{R}_+$ such that the following hold:*

1. *For each $\mu \in (0, \mu^*)$, the full-order system (10) is g.u.e.s. with rate $\alpha_s(\mu)$;*

2. *As $\mu \rightarrow 0$, $\alpha_s(\mu) \rightarrow \alpha_x$ and the gain of exponential convergence for the full-order system remains finite.*

Remark 3. We can summarize the result in Corollary 1 by stating that g.u.e.s. is a robust property for the singularly perturbed systems under consideration in the sense that if it holds for $\mu = 0$, then it holds in a neighborhood of $\mu = 0$. Moreover, the rate of convergence is close to that of the reduced-order system. Theorem 1 adds some more quantitative information. From (21) and (25), we see that the effect on $x(\cdot)$ of a nonzero boundary-layer initial condition y_0 tends to disappear as $\mu \rightarrow 0$. Relationships (22) and (24) demonstrate clearly that the boundary-layer dynamics become faster as μ decreases and, for sufficiently small μ , the “slow” contribution to $y(\cdot)$ is essentially due to nonzero initial x (see (25)). We think it is interesting that the right-hand side of (22) is the sum of an exponential in the τ timescale and an exponential in the t timescale.

Remark 4. Various important results exist in the literature under the generic name of *Tikhonov's theorem*. These results relate the behavior of the full-order system trajectories to those of the reduced-order system and the boundary-layer system (see [10] for a brief description of these results and [13], [7], [8] for details). Roughly speaking, they ensure that, given uniform asymptotic stability of the boundary-layer system and some regularity conditions, then, as $\mu \rightarrow 0$, the solutions of the full-order system converge uniformly to the solutions of the associated complete reduced-order system on any closed time interval *not containing the initial time instant*. The initial time instant can also be included if we consider the solution of the boundary-layer system (see the above-mentioned references). It should be emphasized that (under the hypotheses stated in Assumptions 1–5) this neither implies that the full-order system is g.u.e.s. nor provides the quantitative information given in Theorem 1 and Corollary 1. However, by utilizing Theorem 1, it is possible to prove a version of Tikhonov's theorem.

To this end, let $S \subset \mathbf{R}^n \times \mathbf{R}^m$ be any compact subset of the state space. Denote by $\phi_r(\cdot; t_0, x_0)$ the solution of the reduced-order system (13) corresponding to the initial condition $x(t_0) = x_0$, let $\phi_{y_f}(\cdot; t_0, x_0, y_0)$ be the solution of the boundary-layer system (14) with initial condition $y_f(0) = y_0$, and let

$$\tilde{\phi}_y(\tau; t_0, x_0, y_0) \triangleq \phi_y(t_0 + \mu\tau; t_0, x_0, y_0)$$

for $\mu \neq 0$. Then the following holds.

THEOREM 2 (Tikhonov). *Suppose that Assumptions 1–5 hold and consider any $t_0 \in \mathbf{R}$. Then, as $\mu \rightarrow 0$,*

$$(26) \quad \phi_x(t; t_0, x_0, y_0) \rightarrow \phi_r(t; t_0, x_0),$$

$$(27) \quad \tilde{\phi}_y(\tau; t_0, x_0, y_0) \rightarrow \phi_{y_f}(\tau; t_0, x_0, y_0),$$

uniformly, with respect to $(t, x_0, y_0) \in [t_0, \infty) \times S$ and $(\tau, x_0, y_0) \in [0, \infty) \times S$, respectively.

Proof. Here we will prove only (26). The proof of (27) is left to the interested reader. For the sake of brevity, let $x(t) = \phi_x(t; t_0, x_0, y_0)$, $x_r(t) = \phi_r(t; t_0, x_0)$, $y(t) = \phi_y(t; t_0, x_0, y_0)$. Consider any $\varepsilon > 0$ and define

$$(28) \quad \eta(t) \triangleq x(t) - x_r(t).$$

We show that there exists $\mu_m > 0$ such that for all $\mu \in (0, \mu_m)$, $\|\eta(t)\| < \varepsilon$ for all $(t, x_0, y_0) \in [t_0, \infty) \times S$. From (21)–(25) and Assumption 2, it follows that there exist scalars $\mu_1, \alpha, c > 0$ such that for all $\mu \in (0, \mu_1)$,

$$(29a) \quad \|x(t)\|, \|x_r(t)\| \leq c e^{-\alpha(t-t_0)},$$

$$(29b) \quad \|y(t)\| \leq c e^{-(\alpha/\mu)(t-t_0)} + \mu c,$$

for all $(t, x_0, y_0) \in [t_0, \infty) \times S$.

Since $\|\eta(t)\| \leq \|x(t)\| + \|x_r(t)\|$, it follows from (29a) that there is a $T \geq 0$ such that $\|\eta(t)\| \leq \varepsilon$ for all $t \geq t_0 + T$.

Consider now $t \in [t_0, t_0 + T]$. The evolution of η is governed by the equation

$$(30) \quad \dot{\eta}(t) = F(t, x_r(t) + \eta(t), y(t), \mu) - F(t, x_r(t), 0, 0).$$

Utilizing Assumptions 4 and 5, it is now easy to obtain

$$(31) \quad \|\dot{\eta}\| \leq [M^2 + M + k_f(\mu)]\|\eta\| + k_f(\mu)\|x_r\| + [M + k_f(\mu)]\|y\|,$$

for $\mu \leq \bar{\mu}$, where k_f is continuous with $k_f(0) = 0$. In view of the boundedness of x_r and (29b), there exists $N > 0$ such that

$$(32) \quad \|\dot{\eta}\| \leq N[\|\eta\| + \tilde{k}(\mu) + e^{-(\alpha/\mu)(t-t_0)}],$$

with $\tilde{k}(\mu) \rightarrow 0$ as $\mu \rightarrow 0$. Considering $\eta(t_0) = 0$ and using Gronwall's lemma [5], [2, p. 19], we have that

$$(33) \quad \|\eta(t)\| \leq \tilde{k}(\mu)[e^{N(t-t_0)} - 1] + \mu N(\alpha + \mu N)^{-1}[e^{N(t-t_0)} - e^{-(\alpha/\mu)(t-t_0)}];$$

hence

$$(34) \quad \|\eta(t)\| \leq \tilde{k}(\mu) e^{NT} + \mu N(\alpha + \mu N)^{-1} e^{NT}$$

for $t \in [t_0, t_0 + T]$. Thus there exists $\mu_2 > 0$ such that $\|\eta(t)\| < \varepsilon$ for $t \in [t_0, t_0 + T]$.

Letting $\mu_m = \min\{\mu_1, \mu_2\}$, we have that $\|\eta(t)\| < \varepsilon$ for all $(t, x_0, y_0) \in [t_0, \infty) \times S$. \square

We postpone the proof of Theorem 1 to § 4. There we will use a converse Lyapunov result, which is presented in the next section.

3. A converse Lyapunov result. Converse Lyapunov results are those theorems that, given certain stability properties of a system, ensure the existence of a Lyapunov function that satisfies the Lyapunov conditions for the type of stability under consideration. A thorough description of this class of results can be found in [6]. Our result, which is related to g.u.e.s., is stated in the following lemma.

LEMMA 1. *Consider a parameterized dynamical system*

$$(35) \quad \dot{\xi}(t) = p(t, \xi(t), \theta_1, \theta_2),$$

where $\theta_1 \in \mathbf{R}^{n_{\theta_1}}$ and $\theta_2 \in \mathbf{R}^{n_{\theta_2}}$ are parameter vectors. Suppose that system (35) is g.u.e.s. with rate of convergence α and gain c . In addition, assume that the function p is continuous and continuously differentiable with respect to ξ , θ_1 , θ_2 and that there exist a positive constant γ and functions $k_1, k_2: \mathbf{R}_+ \times \mathbf{R}^{n_{\theta_1}} \times \mathbf{R}^{n_{\theta_2}} \rightarrow \mathbf{R}_+$, continuous and nondecreasing with respect to their first argument, such that

$$(36) \quad \|D_2 p(t, \xi, \theta_1, \theta_2)\| \leq \gamma,$$

$$(37) \quad \|D_3 p(t, \xi, \theta_1, \theta_2)\| \leq k_1(\|\xi\|, \theta_1, \theta_2),$$

$$(38) \quad \|D_4 p(t, \xi, \theta_1, \theta_2)\| \leq k_2(\|\xi\|, \theta_1, \theta_2),$$

for all $t \in \mathbf{R}$, $\xi \in \mathbf{R}^{n_\xi}$, $\theta_1 \in \mathbf{R}^{n_{\theta_1}}$, and $\theta_2 \in \mathbf{R}^{n_{\theta_2}}$.

Then, for any positive $\beta < \alpha$, there exists a parameterized Lyapunov function $V: \mathbf{R} \times \mathbf{R}^{n_\xi} \times \mathbf{R}^{n_{\theta_1}} \times \mathbf{R}^{n_{\theta_2}} \rightarrow \mathbf{R}_+$ satisfying the following inequalities for all $t \in \mathbf{R}$, $\xi \in \mathbf{R}^{n_\xi}$, $\theta_1 \in \mathbf{R}^{n_{\theta_1}}$, and $\theta_2 \in \mathbf{R}^{n_{\theta_2}}$:

$$(39) \quad \omega_1 \|\xi\|^2 \leq V(t, \xi, \theta_1, \theta_2) \leq \omega_2 \|\xi\|^2,$$

$$(40) \quad \begin{aligned} \dot{V}_{(35)}(t, \xi, \theta_1, \theta_2) &\triangleq D_1 V(t, \xi, \theta_1, \theta_2) + D_2 V(t, \xi, \theta_1, \theta_2) p(t, \xi, \theta_1, \theta_2) \\ &\leq -2\beta V(t, \xi, \theta_1, \theta_2), \end{aligned}$$

$$(41a) \quad \|D_2 V(t, \xi, \theta_1, \theta_2)\| \leq \omega_3 \|\xi\|,$$

$$(41b) \quad \|D_3 V(t, \xi, \theta_1, \theta_2)\| \leq \omega_4 k_1(c \|\xi\|, \theta_1, \theta_2) \|\xi\|,$$

$$(41c) \quad \|D_4 V(t, \xi, \theta_1, \theta_2)\| \leq \omega_4 k_2(c \|\xi\|, \theta_1, \theta_2) \|\xi\|,$$

where ω_i , $i = 1, \dots, 4$, are positive constants. Moreover, if system (35) is time-invariant, the Lyapunov function can be chosen to be time-invariant.

Proof. A proof is given at the end of this section.

Remark 5. Note that, using (39) and (40), we can demonstrate that system (35) is g.u.e.s. with rate β and gain $(\omega_2/\omega_1)^{1/2}$, i.e., $\|\xi(t)\| \leq (\omega_2/\omega_1)^{1/2} \|\xi_0\| e^{-\beta(t-t_0)}$.

Remark 6. Although the previous literature contains converse Lyapunov results for exponentially stable systems, we consider the above result to be novel, since it states that we can choose the Lyapunov function so that it recovers arbitrarily closely the information on the rate of convergence of the system.

Remark 7. It should be clear that Lemma 1 also applies for any positive $\beta < \bar{\alpha}$, where $\bar{\alpha}$ is the supremal rate of convergence for (35). In this case, we can select any actual rate of convergence $\alpha \in (\beta, \bar{\alpha})$ and apply Lemma 1. The constant c appearing in inequalities (41) will then be a gain associated with the rate of convergence α .

Remark 8. Sometimes, the supremal rate of convergence $\bar{\alpha}$ is an actual rate of convergence; in this case, Lemma 1 just ensures that it is possible to construct a Lyapunov function satisfying (40) with $\beta < \bar{\alpha}$. However, for linear time-invariant systems, it is possible to show that $\bar{\alpha}$ is an actual rate of convergence *if and only if* there exists a Lyapunov function satisfying (40) with $\beta = \bar{\alpha}$; see Theorem 3.

3.1. Linear time-invariant systems. Lemma 1 can be readily illustrated for linear time-invariant systems. Consider a system described by

$$(42) \quad \dot{\xi} = A\xi,$$

with $\xi \in \mathbf{R}^{n_\xi}$, $A \in \mathbf{R}^{n_\xi \times n_\xi}$, and suppose that

$$(43) \quad -\bar{\alpha} \triangleq \max_{\lambda \in \sigma(A)} \Re(\lambda) < 0.$$

Then system (42) is g.u.e.s. with supremal rate of convergence $\bar{\alpha}$ (see Theorem 3, below). Consider now any $\beta \in (0, \bar{\alpha})$ and any positive definite $Q \in \mathbf{R}^{n_\xi \times n_\xi}$. It can be readily verified that a Lyapunov function V satisfying (39)–(41) is given by $V(\xi) \triangleq \xi^T P \xi$, where P is the unique symmetric positive definite solution of the modified Lyapunov equation

$$P(A + \beta I) + (A + \beta I)^T P + Q = 0.$$

Note that such a solution exists, since the eigenvalues of $A + \beta I$ have negative real parts; see [9].

We also have the following more general result. Denote the eigenvalues of A by λ_i , $i = 1, \dots, m$, where $\lambda_i \neq \lambda_j$ for $i \neq j$. Let \bar{n}_i denote the multiplicity of λ_i , $i = 1, \dots, m$ in the minimal polynomial of A [3]. The following theorem holds.

THEOREM 3. *Consider system (42) and suppose that*

$$-\bar{\alpha} \triangleq \max_{\lambda \in \sigma(A)} \Re(\lambda) < 0.$$

Then (42) is g.u.e.s. with supremal rate of convergence $\bar{\alpha}$, and the following statements are equivalent:

- (i) *The supremal rate of convergence $\bar{\alpha}$ is an actual rate of convergence;*
- (ii) *If $\Re(\lambda_i) = -\bar{\alpha}$, then $\bar{n}_i = 1$, $i = 1, \dots, m$;*
- (iii) *There exists a Lyapunov function $V: \mathbf{R}^{n_\xi} \rightarrow \mathbf{R}_+$ for system (42) such that*

$$\omega_1 \|\xi\|^2 \leq V(\xi) \leq \omega_2 \|\xi\|^2, \quad \dot{V}_{(42)} \leq -2\bar{\alpha} V.$$

Proof. We first prove that, under the given hypothesis, system (42) is g.u.e.s. with supremal rate of convergence $\bar{\alpha}$. To this end, note that the slowest decaying terms in the solutions of (42) have the form

$$(44) \quad t^k e^{-\bar{\alpha}t} \cos(\omega t + \psi) v,$$

where $v \in \mathbf{R}^{n_\xi}$, k is a nonnegative integer, and ω and ψ are real numbers. Hence any number less than $\bar{\alpha}$ is a rate of exponential convergence, whereas any number greater than $\bar{\alpha}$ is not.

Now we prove, by contradiction, that (i) implies (ii). Suppose that there exists an eigenvalue λ_i such that $\Re(\lambda_i) = -\bar{\alpha}$ and $\bar{n}_i > 1$. Then (see [3]) some solutions of (42) will contain terms of the form (44) with $k = 0, 1, \dots, \bar{n}_i - 1$, which, for $k > 0$, cannot be norm-bounded by $c e^{-\bar{\alpha}t}$. In other words, $\bar{\alpha}$ is not an actual rate of convergence; this contradicts (i).

Condition (ii) implies (iii). For the sake of simplicity, let us assume that $\lambda_1 = -\bar{\alpha}$ and $\Re(\lambda_i) < -\bar{\alpha}$ for $i = 2, \dots, m$. Statement (ii) implies that we can consider, without loss of generality,

$$(45) \quad A = \text{blockdiag}(-\bar{\alpha}I_{n_1}, A_R),$$

where $A_R \in \mathbf{R}^{n_R \times n_R}$, $n_R \triangleq n - n_1$, n_1 is the algebraic multiplicity of λ_1 , and $\Re(\lambda) < -\bar{\alpha}$ for all $\lambda \in \sigma(A_R)$; see [3]. Choosing any symmetric positive definite matrix $Q_R \in \mathbf{R}^{n_R \times n_R}$, the modified Lyapunov equation

$$(A_R + \bar{\alpha}I)^T P_R + P_R(A_R + \bar{\alpha}I) + Q_R = 0,$$

has a unique symmetric positive definite solution P_R (see the beginning of this section). We can verify that the sought-after Lyapunov function is

$$V(\xi) \triangleq \xi^T \text{blockdiag}(I_{n_1}, P_R)\xi.$$

We leave to the interested reader the details on the more general case.

The fact that (iii) implies (i) follows from Remark 5. \square

3.2. Proof of Lemma 1. Define the following Lyapunov function candidate:

$$(46) \quad V(t, \xi, \theta_1, \theta_2) \triangleq \int_0^T e^{2\beta\tau} \|\phi(t+\tau; t, \xi, \theta_1, \theta_2)\|^2 d\tau,$$

where T is any positive real number that satisfies

$$(47) \quad T \geq (\alpha - \beta)^{-1} \ln c.$$

We note that if system (35) is time-invariant, then $\phi(t+\tau; t, \xi, \theta_1, \theta_2) = \phi(\tau; 0, \xi, \theta_1, \theta_2)$ and V will not depend on t .

The function V is well defined. Inequality (36) ensures that the function p satisfies a global Lipschitz condition with respect to ξ ; hence a function $\phi(\cdot; t, \xi, \theta_1, \theta_2)$ exists and is unique.

The function V is decrescent. From the hypothesis of g.u.e.s., we have

$$(48) \quad \|\phi(t+\tau; t, \xi, \theta_1, \theta_2)\| \leq c \|\xi\| e^{-\alpha\tau}.$$

Using this inequality, we have that

$$V(t, \xi, \theta_1, \theta_2) \leq c^2 \|\xi\|^2 \int_0^T e^{-2(\alpha-\beta)\tau} d\tau = \omega_2 \|\xi\|^2.$$

The function V is positive definite. We first show that, in view of (36), the norms of the trajectories of system (35) are bounded below as follows:

$$(49) \quad \|\phi(t+\tau; t, \xi, \theta_1, \theta_2)\| \geq \|\xi\| e^{-\gamma\tau} \quad \forall \tau \geq 0.$$

Consider any $t, \xi, \theta_1, \theta_2$ and let $r(\tau) \triangleq -\|\phi(t+\tau; t, \xi, \theta_1, \theta_2)\|^2$. Then

$$(50) \quad \begin{aligned} \dot{r}(\tau) &= -2\phi(t+\tau; t, \xi, \theta_1, \theta_2)^T p(t+\tau, \phi(t+\tau; t, \xi, \theta_1, \theta_2), \theta_1, \theta_2) \\ &\leq 2\gamma \|\phi(t+\tau; t, \xi, \theta_1, \theta_2)\|^2 = -2\gamma r(\tau), \end{aligned}$$

where we have used (36) and the fact that $p(t, 0, \theta_1, \theta_2) \equiv 0$. From (50), we have that $r(\tau) \leq r(0) e^{-2\gamma\tau}$ for all $\tau \geq 0$, and hence (49). Inequality (49) readily leads to

$$V(t, \xi, \theta_1, \theta_2) \geq \omega_1 \|\xi\|^2.$$

The derivative of V along the trajectories of system (35) is negative definite. To prove this, we first obtain the following expression for $V_{(35)}$:

$$\begin{aligned} V(t, \phi(t; t_0, \xi_0, \theta_1, \theta_2), \theta_1, \theta_2) \\ &= \int_0^T e^{2\beta\tau} \|\phi(t+\tau; t, \phi(t; t_0, \xi_0, \theta_1, \theta_2), \theta_1, \theta_2)\|^2 d\tau \\ &= \int_0^T e^{2\beta\tau} \|\phi(t+\tau; t_0, \xi_0, \theta_1, \theta_2)\|^2 d\tau \\ &= e^{-2\beta t} \int_t^{t+T} e^{2\beta\sigma} \|\phi(\sigma; t_0, \xi_0, \theta_1, \theta_2)\|^2 d\sigma. \end{aligned}$$

Then, with $\xi = \phi(t; t_0, \xi_0, \theta_1, \theta_2)$,

$$\begin{aligned} \dot{V}_{(35)}(t, \xi, \theta_1, \theta_2) &= \frac{dV}{dt}(t, \phi(t; t_0, \xi_0, \theta_1, \theta_2), \theta_1, \theta_2) \\ &= -2\beta V(t, \xi, \theta_1, \theta_2) + e^{2\beta T} \|\phi(t+T; t, \xi, \theta_1, \theta_2)\|^2 - \|\xi\|^2 \\ &\leq -2\beta V(t, \xi, \theta_1, \theta_2) + c^2 \|\xi\|^2 e^{-2(\alpha-\beta)T} - \|\xi\|^2, \end{aligned}$$

where we use inequality (48). Considering (47), inequality (40) follows.

The derivatives of V are bounded, as in (41). Since p is continuously differentiable with respect to ξ , θ_1 , and θ_2 , it follows that the function ϕ is continuously differentiable [4, p. 30]. To compute bounds on the derivatives of ϕ with respect to ξ_0 , θ_1 , and θ_2 , note that ϕ satisfies the following integral equation:

$$(51) \quad \phi(t; t_0, \xi_0, \theta_1, \theta_2) = \xi_0 + \int_{t_0}^t p(\tau, \phi(\tau; t_0, \xi_0, \theta_1, \theta_2), \theta_1, \theta_2) d\tau.$$

Let B_{ξ_0} , B_{θ_1} , and B_{θ_2} be compact sets containing ξ_0 , θ_1 , and θ_2 , respectively; then the function $\tilde{p}(\tau, \xi_0, \theta_1, \theta_2) \triangleq p(\tau, \phi(\tau; t_0, \xi_0, \theta_1, \theta_2), \theta_1, \theta_2)$ and its derivatives with respect to ξ_0 , θ_1 , and θ_2 are continuous on $[t_0, t] \times B_{\xi_0} \times B_{\theta_1} \times B_{\theta_2}$. We can then differentiate both sides of (51) with respect to ξ_0 , interchange differentiation and integration, and take norms, obtaining

$$\begin{aligned} (52) \quad &\|D_3\phi(t; t_0, \xi_0, \theta_1, \theta_2)\| \\ &\leq 1 + \int_{t_0}^t \|D_2p(\tau, \phi(\tau; t_0, \xi_0, \theta_1, \theta_2), \theta_1, \theta_2)\| \|D_3\phi(\tau; t_0, \xi_0, \theta_1, \theta_2)\| d\tau \\ &\leq 1 + \int_{t_0}^t \gamma \|D_3\phi(\tau; t_0, \xi_0, \theta_1, \theta_2)\| d\tau. \end{aligned}$$

Applying Gronwall's lemma (see [5]), it follows from (52) that

$$(53) \quad \|D_3\phi(t; t_0, \xi_0, \theta_1, \theta_2)\| \leq e^{\gamma(t-t_0)}.$$

In a similar fashion, we differentiate (51) with respect to θ_i , $i = 1, 2$, interchange differentiation and integration, compute norms, and note that, for $\tau \geq t_0$,

$$\begin{aligned} \|D_{i+2}p(\tau, \phi(\tau; t_0, \xi_0, \theta_1, \theta_2), \theta_1, \theta_2)\| &\leq k_i(\|\phi(\tau; t_0, \xi_0, \theta_1, \theta_2)\|, \theta_1, \theta_2) \\ &\leq k_i(c\|\xi_0\|, \theta_1, \theta_2) \end{aligned}$$

in view of (48) and the hypotheses on the functions k_i . Then Gronwall's lemma yields

$$(54) \quad \|D_{i+3}\phi(t; t_0, \xi_0, \theta_1, \theta_2)\| \leq k_i(c\|\xi_0\|, \theta_1, \theta_2)(t - t_0) e^{\gamma(t-t_0)}, \quad i = 1, 2.$$

Inequalities (53) and (54) hold for all $t_0 \in \mathbf{R}$, $\xi_0 \in \mathbf{R}^{n_\xi}$, $\theta_1 \in \mathbf{R}^{n_{\theta_1}}$, $\theta_2 \in \mathbf{R}^{n_{\theta_2}}$, and $t \geq t_0$.

Now, from

$$(55) \quad \|D_2V(t, \xi, \theta_1, \theta_2)\| \leq \int_0^T 2e^{2\beta\tau} \|\phi(t+\tau; t, \xi, \theta_1, \theta_2)\| \|D_3\phi(t+\tau; t, \xi, \theta_1, \theta_2)\| d\tau,$$

and utilizing inequalities (48) and (53), we have that

$$(56) \quad \|D_2V(t, \xi, \theta_1, \theta_2)\| \leq 2c\|\xi\| \int_0^T e^{(\gamma+2\beta-\alpha)\tau} d\tau = \omega_3\|\xi\|,$$

which results in (41a). The derivations of (41b), (41c) are similar.

4. Proof of Theorem 1. One more lemma is needed.

LEMMA 2. Consider any symmetric matrix $S(\mu)$ given by

$$(57) \quad S(\mu) \triangleq \begin{pmatrix} s_{11}(\mu) & s_{12}(\mu) \\ s_{12}(\mu) & s_{22}(\mu) \end{pmatrix},$$

where the functions s_{11} , s_{12} , $s_{22}: (0, \infty) \rightarrow \mathbf{R}$ satisfy

$$(58) \quad \lim_{\mu \rightarrow 0} s_{11}(\mu) = \lambda_0,$$

$$(59) \quad \lim_{\mu \rightarrow 0} s_{22}(\mu) = \infty,$$

$$(60) \quad \lim_{\mu \rightarrow 0} \frac{s_{12}(\mu)^2}{s_{22}(\mu)} = 0.$$

Then

$$(61) \quad \lim_{\mu \rightarrow 0} \lambda_{\min}[S(\mu)] = \lambda_0.$$

Proof. The characteristic equation for the matrix $S(\mu)$ is given by

$$p^2 - [s_{11}(\mu) + s_{22}(\mu)]p + \Delta(\mu) = 0,$$

where

$$\Delta(\mu) \triangleq s_{11}(\mu)s_{22}(\mu) - s_{12}(\mu)^2.$$

If $\lambda_1(\mu)$, $\lambda_2(\mu)$ denote the eigenvalues of $S(\mu)$, then

$$\begin{aligned} \lambda_1(\mu) &= \frac{s_{11}(\mu) + s_{22}(\mu) - [(s_{11}(\mu) - s_{22}(\mu))^2 + 4s_{12}(\mu)^2]^{1/2}}{2}, \\ \lambda_2(\mu) &= \frac{s_{11}(\mu) + s_{22}(\mu) + [(s_{11}(\mu) - s_{22}(\mu))^2 + 4s_{12}(\mu)^2]^{1/2}}{2}. \end{aligned}$$

Since $\Delta(\mu) = \lambda_1(\mu)\lambda_2(\mu)$ and, for sufficiently small μ , $\lambda_2(\mu) > 0$ and $s_{22}(\mu) > 0$, we have that

$$(62) \quad \begin{aligned} \lambda_{\min}[S(\mu)] &= \lambda_1 = \Delta/\lambda_2 \\ &= \frac{2(s_{11} - s_{12}^2/s_{22})}{1 + s_{11}/s_{22} + [(s_{11}/s_{22} - 1)^2 + 4s_{12}^2/s_{22}^2]^{1/2}}. \end{aligned}$$

Taking the limit as $\mu \rightarrow \infty$ in (62) and considering the hypotheses, the desired result follows. \square

We can now proceed with the proof of Theorem 1.

Proof of Theorem 1. First, we prove that the functions F and G in (10) possess the same qualitative properties as those prescribed for f and g in Assumptions 4 and 5.

From definitions (11), (12), and Assumption 4, it is readily seen that

$$(63a) \quad \|D_2 F(t, x, y, 0)\| \leq M + M^2,$$

$$(63b) \quad \|D_3 F(t, x, y, 0)\| \leq M,$$

$$(63c) \quad \|D_1 G(t, x, y, 0)\| \leq (M + M^2)\|x\| + M\|y\|,$$

$$(63d) \quad \|D_2 G(t, x, y, 0)\| \leq M + M^2,$$

$$(63e) \quad \|D_3 G(t, x, y, 0)\| \leq M,$$

for all $t \in \mathbf{R}$, $x \in \mathbf{R}^n$ and $y \in \mathbf{R}^m$. From Assumption 5, we have that

$$(64) \quad \|F(t, x, y, \mu) - F(t, x, y, 0)\| \leq k_f(\mu)(\|x\| + \|y\|).$$

From this and (63), it follows that

$$(65) \quad \|F(t, x, y, \mu)\| \leq [M + M^2 + k_f(\mu)]\|x\| + [M + k_f(\mu)]\|y\|.$$

Using Assumption 5 and inequalities (63), (65), we can show that

$$(66) \quad \|G(t, x, y, \mu) - G(t, x, y, 0)\| \leq k_G(\mu)(\|x\| + \|y\|),$$

where $k_G: [0, \bar{\mu}] \rightarrow \mathbf{R}_+$ is a continuous function such that $k_G(0) = 0$, and there exists $d_G \geq 0$ with $k_G(\mu)/\mu \leq d_G$ for $\mu \in (0, \bar{\mu})$.

Since the reduced-order system (3) is g.u.e.s. with rate of convergence $\alpha_x < \bar{\alpha}_x$ and it satisfies the hypotheses of Lemma 1, it follows that (see Remark 7) there exists a Lyapunov function $V: \mathbf{R} \times \mathbf{R}^n \rightarrow \mathbf{R}_+$ for system (3) such that

$$(67) \quad \omega_1 \|x\|^2 \leq V(t, x) \leq \omega_2 \|x\|^2,$$

$$(68) \quad \dot{V}_{(3)}(t, x) \leq -2\alpha_x V(t, x),$$

$$(69) \quad \|D_2 V(t, x)\| \leq \omega_3 \|x\|,$$

for some positive constants ω_i , $i = 1, 2, 3$.

Similarly, the boundary-layer system (8) is g.u.e.s. with rate of convergence $\alpha_y < \bar{\alpha}_y$, and it satisfies the hypotheses of Lemma 1; hence there exists a Lyapunov function $W: \mathbf{R} \times \mathbf{R}^n \times \mathbf{R}^m \rightarrow \mathbf{R}_+$ such that

$$(70) \quad \nu_1 \|y_f\|^2 \leq W(t_0, x_0, y_f) \leq \nu_2 \|y_f\|^2,$$

$$(71) \quad \dot{W}_{(8)}(t_0, x_0, y_f) \leq -2\alpha_y W(t_0, x_0, y_f),$$

$$(72) \quad \|D_1 W(t_0, x_0, y_f)\| \leq \nu_3 \|y_f\|^2 + \nu_3 \|x_0\| \|y_f\|,$$

$$(73) \quad \|D_2 W(t_0, x_0, y_f)\| \leq \nu_3 \|y_f\|,$$

$$(74) \quad \|D_3 W(t_0, x_0, y_f)\| \leq \nu_3 \|y_f\|,$$

for positive constants ν_i , $i = 1, 2, 3$.

The derivative of V along trajectories of the full-order system (10) is given by

$$(75) \quad \begin{aligned} \dot{V}_{(10)}(t, x, y, \mu) = & [D_1 V(t, x) + D_2 V(t, x)F(t, x, 0, 0)] \\ & + D_2 V(t, x)[F(t, x, y, \mu) - F(t, x, 0, 0)]. \end{aligned}$$

From (63) and (64),

$$\begin{aligned} \|F(t, x, y, \mu) - F(t, x, 0, 0)\| &\leq \|F(t, x, y, \mu) - F(t, x, y, 0)\| \\ &\quad + \|F(t, x, y, 0) - F(t, x, 0, 0)\| \\ &\leq k_f(\mu)\|x\| + [M + k_f(\mu)]\|y\|; \end{aligned}$$

hence, using (68) and (69), (75) yields

$$(76) \quad \dot{V}_{(10)} \leq -2\alpha_x V + \omega_3 \|x\| [k_f \|x\| + (M + k_f)\|y\|].$$

Finally, considering (67) and (70), we obtain

$$(77) \quad \dot{V}_{(10)} \leq -2s_{11}(\mu)V + 2s_{12}(\mu)V^{1/2}W^{1/2},$$

where

$$(78) \quad s_{11}(\mu) \triangleq \alpha_x - k_f(\mu)\omega_3(2\omega_1)^{-1},$$

$$(79) \quad s_{12}(\mu) \triangleq \omega_3[M + k_f(\mu)](4\omega_1\nu_1)^{-1/2},$$

and the arguments of W are t, x, y .

The derivative of $W(\cdot, x(\cdot), y(\cdot))$ along a trajectory of system (10) satisfies

$$(80) \quad \begin{aligned} \dot{W}_{(10)}(t, x, y, \mu) &\leq \|D_1 W(t, x, y)\| + \|D_2 W(t, x, y)\| \|F(t, x, y, \mu)\| \\ &\quad + \mu^{-1} D_3 W(t, x, y) G(t, x, y, 0) \\ &\quad + \mu^{-1} \|D_3 W(t, x, y)\| \|G(t, x, y, \mu) - G(t, x, y, 0)\|. \end{aligned}$$

Using inequalities (65), (66), (71)–(74), we obtain

$$(81) \quad \begin{aligned} \dot{W}_{(10)} &\leq -2\alpha_y \mu^{-1} W + \nu_3(1 + M + M^2 + d_G + k_f)\|x\| \|y\| \\ &\quad + \nu_3(1 + M + d_G + k_f)\|y\|^2, \end{aligned}$$

which holds for all $\mu \in (0, \bar{\mu})$. Again utilizing (67) and (70), the last inequality yields

$$(82) \quad \dot{W}_{(10)} \leq 2s_{21}(\mu)V^{1/2}W^{1/2} - 2s_{22}(\mu)W,$$

where

$$(83) \quad s_{21}(\mu) \triangleq \nu_3[1 + M + M^2 + d_G + k_f(\mu)](4\omega_1\nu_1)^{-1/2},$$

$$(84) \quad s_{22}(\mu) \triangleq \alpha_y \mu^{-1} - \nu_3[1 + M + d_G + k_f(\mu)](2\nu_1)^{-1}.$$

Now consider the following Lyapunov function candidate for (10):

$$(85) \quad L(t, x, y) \triangleq V(t, x) + \kappa(\mu)W(t, x, y),$$

where $\kappa : (0, \bar{\mu}) \rightarrow \mathbf{R}_+$ is any continuous function that satisfies

$$(86a) \quad \lim_{\mu \rightarrow 0} \kappa(\mu) = 0,$$

$$(86b) \quad \lim_{\mu \rightarrow 0} \mu/\kappa(\mu) = 0;$$

e.g., consider $\kappa(\mu) = \mu^{1/2}$.

It is readily seen that L satisfies the inequalities

$$(87a) \quad L(t, x, y) \geq \omega_1 \|x\|^2 + \kappa(\mu) \nu_1 \|y\|^2,$$

$$(87b) \quad L(t, x, y) \leq \omega_2 \|x\|^2 + \kappa(\mu) \nu_2 \|y\|^2.$$

In view of (77) and (82), the derivative of L along trajectories of system (10) has the following bound:

$$(88) \quad \dot{L}_{(10)} \leq -(V^{1/2} \kappa(\mu)^{1/2} W^{1/2}) S(\mu) \begin{pmatrix} V^{1/2} \\ \kappa(\mu)^{1/2} W^{1/2} \end{pmatrix},$$

where

$$(89) \quad S(\mu) \triangleq \begin{pmatrix} 2s_{11}(\mu) & s'_{12}(\mu) \\ s'_{12}(\mu) & 2s_{22}(\mu) \end{pmatrix},$$

$$(90) \quad s'_{12}(\mu) \triangleq -\kappa(\mu)^{-1/2} s_{12}(\mu) - \kappa(\mu)^{1/2} s_{21}(\mu).$$

Note that, from inequality (88), we can obtain

$$(91) \quad \dot{L}_{(10)} \leq -\lambda_{\min}[S(\mu)]L.$$

Since

$$\lim_{\mu \rightarrow 0} 2s_{11}(\mu) = 2\alpha_x, \quad \lim_{\mu \rightarrow 0} 2s_{22}(\mu) = \infty,$$

$$\lim_{\mu \rightarrow 0} s'_{12}(\mu)^2 / (2s_{22}(\mu)) = 0,$$

the matrix function $S(\cdot)$ in (91) satisfies the hypotheses of Lemma 2; hence, defining

$$(92) \quad \alpha_s(\mu) \triangleq \lambda_{\min}[S(\mu)]/2,$$

we have that

$$(93) \quad \lim_{\mu \rightarrow 0} \alpha_s(\mu) = \alpha_x.$$

From the continuity of the function $\alpha_s(\cdot)$, there exists $\mu_1^* > 0$ with $\mu_1^* \leq \bar{\mu}$, such that $\alpha_s(\mu) > 0$ for all $\mu \in (0, \mu_1^*)$.

From (91), (92), it follows that

$$(94) \quad L_{(10)}(t) \leq L_{(10)}(t_0) e^{-2\alpha_s(\mu)(t-t_0)}$$

and, since $V(t, x) \leq L(t, x, y)$ for all $t \in \mathbf{R}$, $x \in \mathbf{R}^n$, $y \in \mathbf{R}^m$,

$$(95) \quad V_{(10)}(t) \leq [\omega_2 \|x_0\|^2 + \nu_2 \kappa(\mu) \|y_0\|^2] e^{-2\alpha_s(\mu)(t-t_0)},$$

and we have used inequality (87b). Applying (67) to (95) and recalling that $a^2 + b^2 \leq (a+b)^2$ for any $a, b \geq 0$ leads to

$$(96) \quad \|x(t)\| \leq c_2 [\|x_0\| + \gamma(\mu) \|y_0\|] e^{-\alpha_s(\mu)(t-t_0)},$$

where

$$(97) \quad c_2 \triangleq (\omega_2/\omega_1)^{1/2}, \quad \gamma(\mu) \triangleq (\nu_2 \kappa(\mu)/\omega_2)^{1/2}.$$

Recalling (82) and using inequality (95) yields

$$(98) \quad \dot{W}_{(10)} \leq 2c_V e^{-\alpha_s(\mu)(t-t_0)} W^{1/2} - 2s_{22} W,$$

where

$$(99) \quad c_V \triangleq s_{21} [\omega_2^{1/2} \|x_0\| + (\nu_2 \kappa)^{1/2} \|y_0\|].$$

Let $w(t) \triangleq [W_{(10)}(t)]^{1/2}$. Note that w is not necessarily differentiable when $w(t) = 0$. For $w(t) > 0$,

$$(100) \quad \dot{w} \leq -s_{22}w + c_V e^{-\alpha_s(t-t_0)}.$$

Let $\bar{w}(\cdot)$ be the solution of the differential equation

$$(101) \quad \dot{\bar{w}} = -s_{22}\bar{w} + c_V e^{-\alpha_s(t-t_0)}$$

with $\bar{w}(t_0) = w(t_0) \geq 0$.

We now show that

$$(102) \quad w(t) \leq \bar{w}(t) \quad \forall t \geq t_0.$$

First, we note that $\bar{w}(t_0) \geq 0$ implies that $\bar{w}(t) \geq 0$ for all $t \geq t_0$. Suppose now that there exists $t_1 > t_0$ such that $w(t_1) > \bar{w}(t_1)$ and define $\eta(t) \triangleq w(t) - \bar{w}(t)$. Then $\eta(t_1) > 0$ and, since $\eta(t_0) = 0$ and η is continuous, there exists a $t'_0 \in [t_0, t_1)$ such that $\eta(t'_0) = 0$ and $\eta(t) > 0$ for $t \in (t'_0, t_1)$. This implies that η is differentiable in (t'_0, t_1) and its derivative satisfies

$$\dot{\eta} = \dot{w} - \dot{\bar{w}} \leq -s_{22}(\mu)\eta < 0.$$

Since the mean value theorem [1, p. 196] requires the existence of a point $\bar{t} \in (t'_0, t_1)$, where

$$\dot{\eta}(\bar{t}) = \frac{\eta(t_1) - \eta(t'_0)}{t_1 - t'_0} > 0,$$

the required result (102) follows by contradiction.

For sufficiently small μ , say $\mu \leq \mu_2^*$, $s_{22}(\mu) - \alpha_s(\mu) > 0$; hence, for $0 \leq \mu \leq \mu^*$ where $\mu^* \triangleq \min\{\mu_1^*, \mu_2^*\}$, we obtain from (102) and (101)

$$(103) \quad w(t) \leq \bar{w}(t) = w(t_0) e^{-s_{22}(t-t_0)} + c_V(s_{22} - \alpha_s)^{-1} e^{-\alpha_s(t-t_0)}.$$

Finally, from (70),

$$(104) \quad \|y(t)\| \leq c_3 \|y_0\| e^{-\alpha_f(\mu)(t-t_0)/\mu} + \mu c_4 [\|x_0\| + \gamma(\mu)\|y_0\|] e^{-\alpha_s(\mu)(t-t_0)},$$

where

$$(105) \quad \alpha_f(\mu) \triangleq \mu s_{22}(\mu),$$

$$(106) \quad c_3 \triangleq (\nu_2/\nu_1)^{1/2},$$

$$(107) \quad c_4 \triangleq \sup \left\{ \frac{s_{21}(\mu)(\omega_2)^{1/2}}{\nu_1^{1/2}[\alpha_f(\mu) - \mu\alpha_s(\mu)]} : \mu \in [0, \mu^*] \right\}.$$

Letting

$$(108) \quad c_1 \triangleq \max\{c_2, c_3, c_4\},$$

inequalities (21), (22) of Theorem 1 follow from (96), (104), respectively. Limit (25) follows from (97), (86a); also, (24) follows from (105), (84). \square

Remark 9. The proof of Theorem 1 is based on the existence of Lyapunov functions V and W , which satisfy (67)–(69) and (70)–(74), respectively; this existence is ensured by Lemma 1, provided that $\alpha_x < \bar{\alpha}_x$ and $\alpha_y < \bar{\alpha}_y$. However, as pointed out in Remark 8, it may occur that for the reduced-order system and/or the boundary-layer system, the supremal rate of convergence is an actual rate of convergence. Suppose that $\bar{\alpha}_x$ is a rate of convergence. Then if a function V exists that satisfies (67)–(69) with $\alpha_x = \bar{\alpha}_x$, Theorem 1 holds, with $\bar{\alpha}_x$ replacing α_x in (23).

5. Two examples.

5.1. First example. Consider the following system:

$$(109a) \quad \dot{x} = -z,$$

$$(109b) \quad \mu \dot{z} = x - z,$$

where $x, z \in \mathbf{R}$. Setting $\mu = 0$, we obtain $z \triangleq h(t, x) = x$ and, after defining $y \triangleq z - x$, we can rewrite the above system in the following form:

$$(110a) \quad \dot{x} = -x - y,$$

$$(110b) \quad \mu \dot{y} = \mu x - (1 - \mu)y.$$

The reduced-order and boundary-layer systems are

$$(111) \quad \dot{x} = -x,$$

$$(112) \quad \frac{dy_f}{d\tau} = -y_f,$$

respectively, where the subscript f in (112) emphasizes the change of timescale. We note that they have the same dynamics and, in particular, that they are g.u.e.s. with supremal rate of convergence equal to 1. Moreover, since condition (ii) of Theorem 3 holds for both systems, they admit Lyapunov functions given by $V(t, x) \triangleq x^2$ and $W(t_0, x_0, y_f) = y_f^2$, which satisfy (67)–(69) and (70)–(74) with $\alpha_x = \alpha_y = 1$. Hence (110) has the properties given by Theorem 1 with the modification suggested in Remark 9.

To directly verify our assertions, consider that the solution of (110) with initial conditions $x(0) = x_0$, $y(0) = y_0$ is given by

$$(113a) \quad \begin{aligned} x(t) = & [(1/2)(1+a)x_0 - \mu by_0] e^{\lambda_s(\mu)t} \\ & + [(1/2)(1-a)x_0 + \mu by_0] e^{\lambda_f(\mu)t}, \end{aligned}$$

$$(113b) \quad \begin{aligned} y(t) = & [\mu bx_0 + (1/2)(1-a)y_0] e^{\lambda_s(\mu)t} \\ & + [-\mu bx_0 + (1/2)(1+a)y_0] e^{\lambda_f(\mu)t}, \end{aligned}$$

where

$$a(\mu) \triangleq (1 - 2\mu)(1 - 4\mu)^{-1/2},$$

$$b(\mu) \triangleq (1 - 4\mu)^{-1/2},$$

$$\lambda_s(\mu) \triangleq -2[1 + \sqrt{1 - 4\mu}]^{-1} \triangleq -\alpha_s(\mu),$$

$$\lambda_f(\mu) \triangleq -(2\mu)^{-1}[(1 + \sqrt{1 - 4\mu})] \triangleq -\mu^{-1}\alpha_f(\mu),$$

and we consider $\mu < \frac{1}{4}$.

Now, as $\mu \rightarrow 0$,

$$(114) \quad \alpha_s(\mu) \triangleq -\lambda_s(\mu) \rightarrow -1, \quad \alpha_f(\mu) \triangleq -\mu\lambda_f(\mu) \rightarrow -1.$$

Since, for sufficiently small μ ,

$$(115a) \quad a(\mu) - 1 \cong 2\mu^2,$$

$$(115b) \quad b(\mu) \leq 2,$$

$$(115c) \quad \lambda_f(\mu) < \lambda_s(\mu),$$

$$(115d) \quad (1 + a(\mu)) e^{\lambda_f(\mu)t} < 2e^{\lambda_f(\mu)t} + (a(\mu) - 1) e^{\lambda_s(\mu)t} \quad \forall t,$$

we readily obtain

$$(116a) \quad \begin{aligned} |x(t)| &\leq [|x_0| + \mu 2b|y_0|] e^{-\alpha_s(\mu)t} \\ &\leq 4[|x_0| + 2\mu|y_0|] e^{-\alpha_s(\mu)t}, \end{aligned}$$

$$(116b) \quad \begin{aligned} |y(t)| &\leq |y_0| e^{-\alpha_f(\mu)t/\mu} + \mu[2b|x_0| + \mu^{-1}(a-1)|y_0|] e^{-\alpha_s(\mu)t} \\ &\leq 4|y_0| e^{-\alpha_f(\mu)t/\mu} + 4\mu[|x_0| + 2\mu|y_0|] e^{-\alpha_s(\mu)t}, \end{aligned}$$

i.e., inequalities (21), (22) of Theorem 1. The properties in statement 2 of Theorem 1 are also satisfied.

5.2. Feedback stabilization of a nonlinear flexible mechanical system. Consider the mechanical system illustrated in Fig. 1. There an inverted pendulum, consisting of a uniform bar of length $2l$ and mass m , is connected to one end of a massless shaft. At the other end of the shaft is a rotor to which a control torque u is applied. We are interested in the following problem. Suppose, for simplicity in design, we model the shaft as rigid and design a linear feedback controller to render the closed-loop rigid model g.u.e.s. Will the same controller stabilize a model in which the shaft is flexible, but sufficiently stiff?

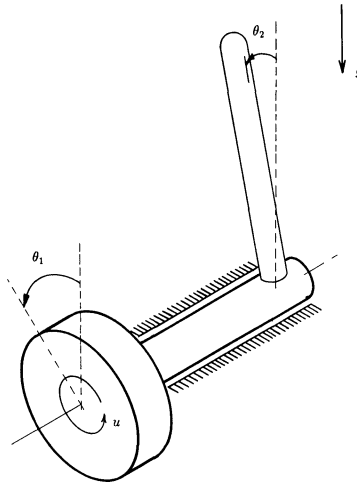


FIG. 1. A mechanical system.

With the shaft rigid, the angular displacements θ_1 and θ_2 are always equal, i.e., $\theta_1 = \theta_2$, and the motion of the system is described by

$$(117) \quad (I_1 + I_2)\ddot{\theta}_2 = mgl \sin \theta_2 + u,$$

where I_1 is the moment of inertia of the rotor, I_2 is the moment of inertia of the pendulum, and g is the (lunar) gravitational acceleration constant. Suppose that θ_2 and $\dot{\theta}_2$ can be measured. Then a linear controller that guarantees g.u.e.s. of system (117) is given by

$$(118) \quad u = -k_1\theta_2 - k_2\dot{\theta}_2,$$

provided that

$$(119) \quad k_1 > mgl, \quad k_2 > 0.$$

To see this, consider the closed-loop system

$$(120) \quad (I_1 + I_2)\ddot{\theta}_2 = mgl \sin \theta_2 - k_1\theta_2 - k_2\dot{\theta}_2;$$

now define

$$(121) \quad x_1 \triangleq \theta_2, \quad x_2 \triangleq \dot{\theta}_2$$

and rewrite (120) as follows:

$$(122a) \quad \dot{x}_1 = x_2,$$

$$(122b) \quad \dot{x}_2 = -I_T^{-1}(k_1x_1 - mgl \sin x_1 + k_2x_2),$$

with $I_T \triangleq I_1 + I_2$. g.u.e.s. of (122) can be shown using the Lyapunov function $V(x) \triangleq x^T Px + 2mgl(\cos x_1 - 1)$, with

$$P = \begin{pmatrix} k_1 + \rho k_2^2 / I_T & \rho k_2 \\ \rho k_2 & I_T \end{pmatrix}, \quad \rho \in (0, 1).$$

Suppose now that the shaft is not rigid, but is modelled as a parallel combination of a linear torsional spring of spring constant $\beta_s \mu^{-2} > 0$ and a linear torsional damper of damping coefficient $\beta_d \mu^{-1} > 0$. Then, in general, $\theta_1 \neq \theta_2$, and a description of the system is

$$(123a) \quad I_1 \ddot{\theta}_1 + I_2 \ddot{\theta}_2 = mgl \sin \theta_2 + u,$$

$$(123b) \quad I_2 \ddot{\theta}_2 = -\beta_d \mu^{-1}(\dot{\theta}_2 - \dot{\theta}_1) - \beta_s \mu^{-2}(\theta_2 - \theta_1) + mgl \sin \theta_2.$$

To determine the stability of (123) with the feedback law (118), we could rewrite the closed-loop equations as a system of four first-order ordinary differential equations and then analyze it with usual tools, for example, Lyapunov's second method. Alternatively, we could rewrite (123), (118) as a singularly perturbed system in the standard form (1) and then show (hopefully) that both the reduced-order system and the boundary-layer system associated with that representation are g.u.e.s.

Taking the latter approach, let

$$(124) \quad z_1 \triangleq \mu^{-2}(\theta_2 - \theta_1), \quad z_2 \triangleq \mu^{-1}(\dot{\theta}_2 - \dot{\theta}_1).$$

Equations (121), (124), (123), and (118) readily yield

$$(125a) \quad \dot{x}_1 = x_2,$$

$$(125b) \quad \dot{x}_2 = I_2^{-1}(mgl \sin x_1 - \beta_s z_1 - \beta_d z_2),$$

$$(125c) \quad \mu \dot{z}_1 = z_2,$$

$$(125d) \quad \mu \dot{z}_2 = I_1^{-1}(k_1x_1 + k_2x_2) + I_2^{-1}mgl \sin x_1 - I_P^{-1}(\beta_s z_1 + \beta_d z_2),$$

with $I_P \triangleq I_1 I_2 (I_1 + I_2)^{-1}$.

Letting $\mu = 0$, we obtain

$$(126a) \quad z_1 = h_1(t, x) = (\beta_s I_T)^{-1}[I_2(k_1x_1 + k_2x_2) + I_1 mgl \sin x_1],$$

$$(126b) \quad z_2 = h_2(t, x) = 0,$$

which, upon substitution into (125a), (125b) yields (122), i.e., the closed-loop rigid model. Hence the reduced-order system is g.u.e.s.

Note that, on rewriting (123b) as

$$\mu^2 I_2 \ddot{\theta}_2 = -\beta_d \mu(\dot{\theta}_2 - \dot{\theta}_1) - \beta_s(\theta_2 - \theta_1) + \mu^2 mgl \sin \theta_2$$

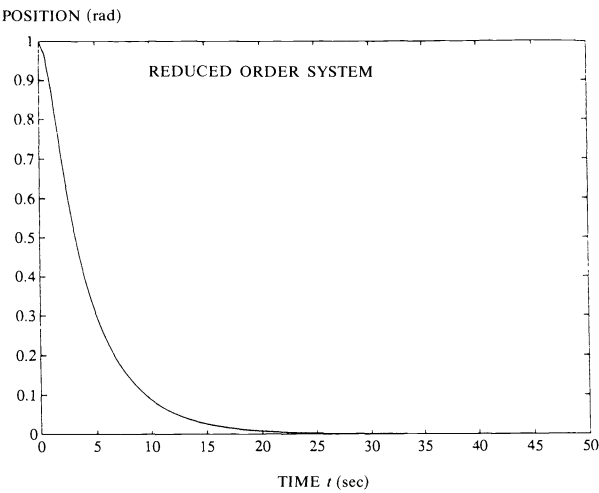


FIG. 2. Time history of the angular position of the pendulum for the reduced-order system.

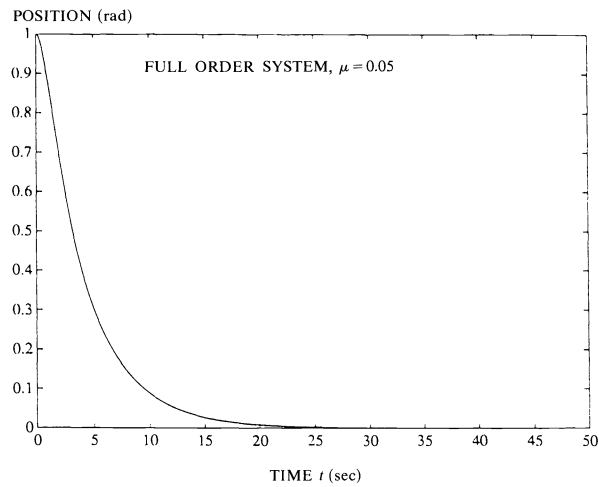


FIG. 3. Time history of the angular position of the pendulum for the full-order system with $\mu = 0.05$.

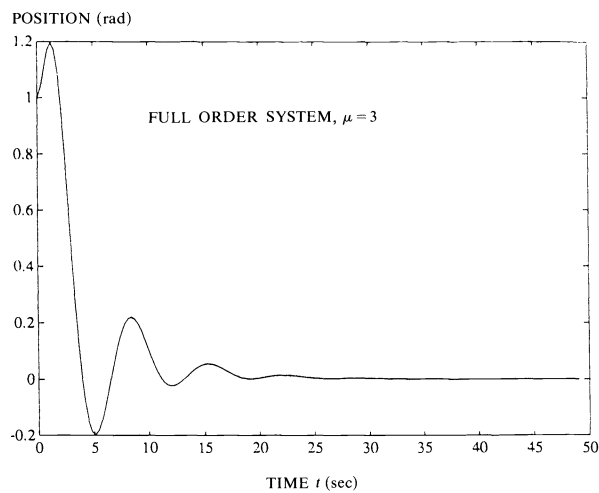


FIG. 4. Time history of the angular position of the pendulum for the full-order system with $\mu = 3$.

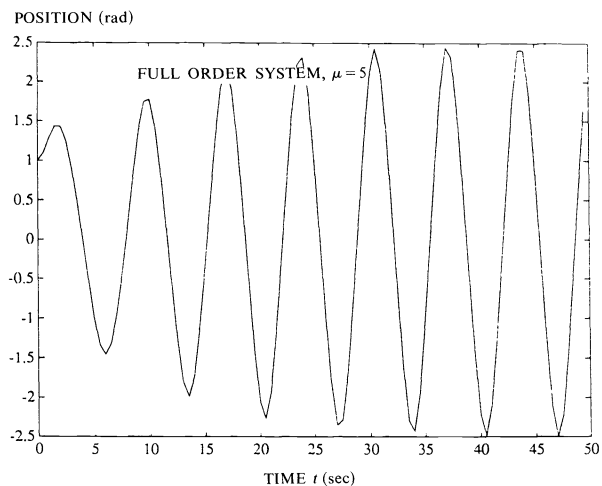


FIG. 5. Time history of the angular position of the pendulum for the full-order system with $\mu = 5$.

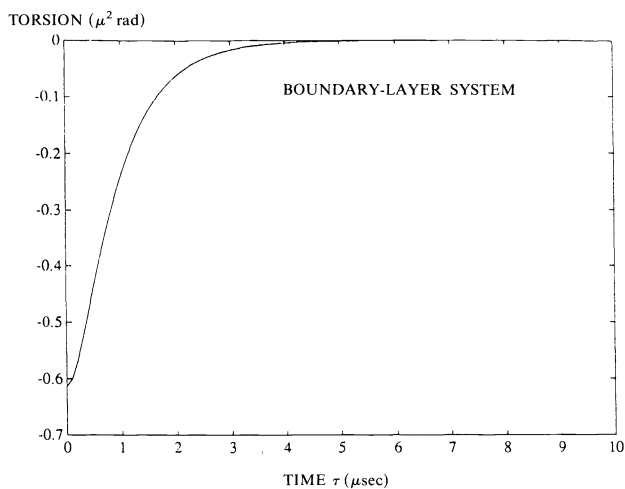


FIG. 6. Time history of the variable y_1 for the boundary-layer system.

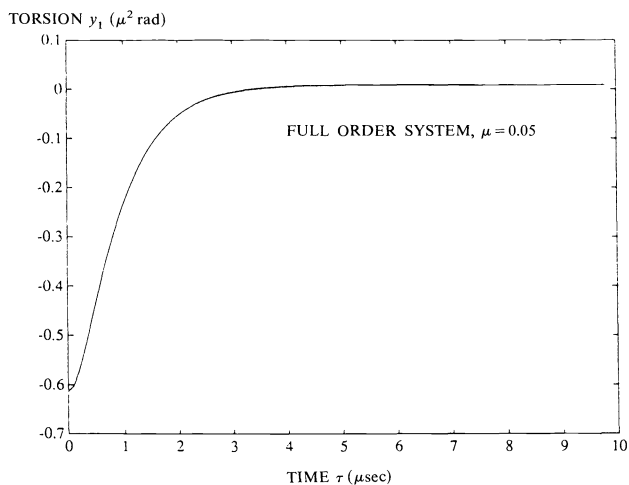


FIG. 7. Time history of the variable y_1 for the full-order system with $\mu = 0.05$. The time unit is μsec .

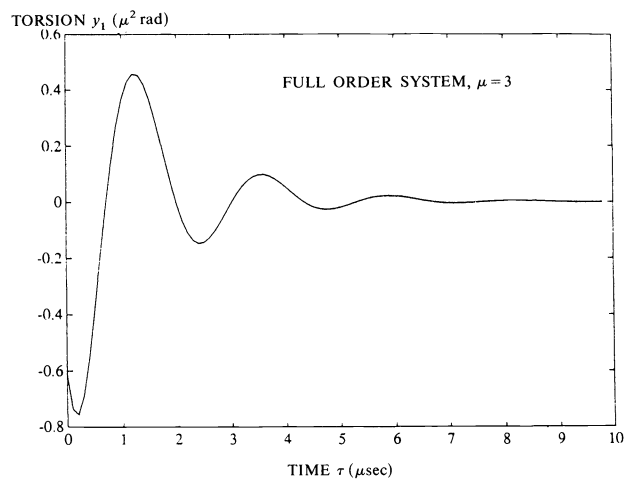


FIG. 8. Time history of the variable y_1 for the full-order system with $\mu = 3$. The time unit is μsec .

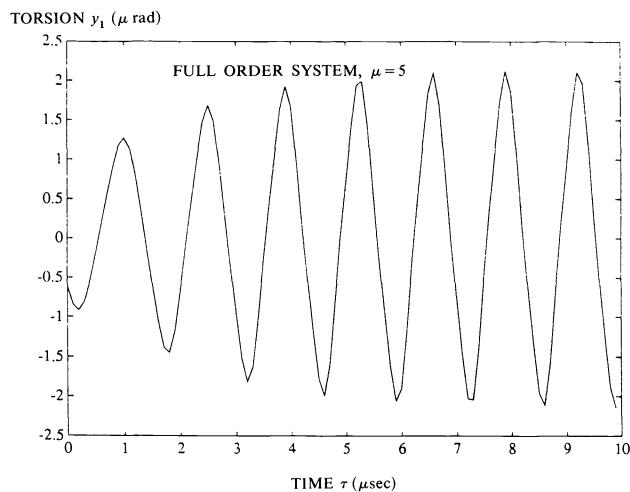


FIG. 9. Time history of the variable y_1 for the full-order system with $\mu = 5$. The time unit is μsec .

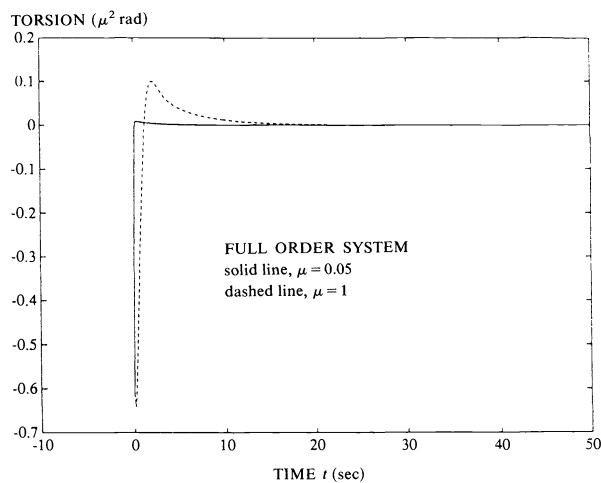


FIG. 10. Time history of the variable y_1 for the full-order system with $\mu = 0.05$ and $\mu = 1$. The time unit is sec.

and letting $\mu = 0$, we obtain $\theta_1 \equiv \theta_2$. Substituting this into (123a) and utilizing (118) yields system (120).

Defining

$$(127) \quad y_1 \triangleq z_1 - h_1(t, x), \quad y_2 \triangleq z_2 - h_2(t, x),$$

the boundary-layer systems is given by

$$(128a) \quad \frac{dy_{1f}}{d\tau} = y_{2f},$$

$$(128b) \quad \frac{dy_{2f}}{d\tau} = -I_P^{-1}(\beta_s y_{1f} + \beta_d y_{2f}),$$

which is g.u.e.s. In (128) we use the subscript f to emphasize the change of timescale.

It now follows from Theorem 1 that the closed-loop flexible model is g.u.e.s., provided that the shaft is stiff enough, i.e., provided that $\mu > 0$ is sufficiently small.

Numerical simulation results. Here we present some numerical simulation results performed with

$$I_1 = 1 \text{ kg m}^2, \quad m = 1 \text{ kg}, \quad l = 1 \text{ m}, \\ g = 1.62 \text{ m sec}^{-2}, \quad \beta_s = 3 \text{ N m}, \quad \beta_d = 3 \text{ N m sec}.$$

The control parameters are $k_1 = 2.2$, $k_2 = 3$. The initial conditions are

$$\theta_1(0) = \theta_2(0) = 1 \text{ rad}, \quad \dot{\theta}_1(0) = \dot{\theta}_2(0) = 0 \text{ rad sec}^{-1}.$$

Figure 2 illustrates the behavior of the reduced-order system; there the angular position of the pendulum, i.e., the variable x_1 , is plotted against time t . Figures 3 and 4 show that, for increasing values of the parameter μ , the closed-loop flexible model remains exponentially stable, although with decreasing rates of convergence. In particular, no appreciable difference exists between the plots in Figs. 2 and 3. In Fig. 5 we see that the flexible model is unstable for $\mu = 5$. Similar comments apply to Figs. 6–9, in which the boundary-layer variable y_{1f} is plotted against the fast time variable τ . Since the unit of time in these plots is μsec , the duration of the boundary-layer phenomena changes with μ . This is highlighted in Fig. 10, where the variable y_1 is plotted against time t for two different values of μ .

6. Conclusions. In this paper we have presented new results on the exponential stability of singularly perturbed systems. In particular, we have seen that, if the reduced-order and boundary-layer systems are globally uniformly exponentially stable, then, provided that some further regularity conditions are satisfied, it is possible to establish exponential bounds on the norms of the trajectories of the full-order system. These exponential bounds depend on the “slow” time variable t and on the “fast” time variable τ and have rates of convergence that, as $\mu \rightarrow 0$, are arbitrarily close to those of the reduced-order and boundary-layer systems. A useful converse Lyapunov result for exponentially stable systems has also been presented.

Acknowledgment. The authors thank an anonymous reviewer for Remark 1.

REFERENCES

- [1] R. G. BARTLE, *The Elements of Real Analysis*, 2nd ed., John Wiley, New York, 1976.
- [2] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.
- [3] C. T. CHEN, *Introduction to Linear System Theory*, Holt, Rinehart and Winston, New York, 1970.
- [4] E. A. CODDINGTON AND N. LEVINSON, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

- [5] T. H. GRONWALL, *Note on the derivatives with respect to a parameter of the solutions of a system of differential equations*, Ann. Math., 20 (1918), pp. 292–296.
- [6] W. HAHN, *Stability of Motion*, Springer-Verlag, Berlin, 1967.
- [7] F. HOPPENSTEADT, *Singular perturbations on the infinite interval*, Trans. Amer. Math. Soc., 123 (1966), pp. 521–535.
- [8] ———, *Properties of solutions of ordinary differential equations with small parameters*, Comm. Pure Appl. Math., 24 (1971), pp. 807–840.
- [9] R. E. KALMAN AND J. E. BERTRAM, *Control system analysis and design via the “second method” of Lyapunov, I: Continuous-time systems*, ASME J. Basic Engrg., 82 (1960), pp. 371–393.
- [10] P. V. KOKOTOVIĆ, H. K. KHALIL, AND J. O'REILLY, *Singular Perturbation Methods in Control: Analysis and Design*, Academic Press, London, 1986.
- [11] P. V. KOKOTOVIĆ AND H. K. KHALIL, EDS. *Singular Perturbations in Systems and Control*, IEEE Press, New York, 1986.
- [12] A. SABERI AND H. KHALIL, *Quadratic-type Lyapunov functions for singularly perturbed systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 542–550.
- [13] A. N. TIKHONOV, *Systems of differential equations containing a small parameter multiplying the highest derivatives*, Mat. Sb., NS(31) 73 (1952), pp. 575–585. (In Russian.)

VARIANTS OF THE KUHN–TUCKER SUFFICIENT CONDITIONS IN CONES OF NONNEGATIVE FUNCTIONS*

J. C. DUNN† AND T. TIAN†

Abstract. Second-order sufficient conditions of the Kuhn–Tucker type are proved for certain constrained minimization problems on sets of nonnegative \mathcal{L}^p functions, with $p \in [2, \infty]$. The objective functions for these problems have specially structured bilinear second Gateaux differentials that are bounded with respect to the \mathcal{L}^2 norm and vary continuously with respect to the \mathcal{L}^2 norm on \mathcal{L}^p . Structure and smoothness conditions of this sort are satisfied by nontrivial classes of constrained-input Bolza optimal control problems, and in this context, the associated Kuhn–Tucker sufficient conditions yield a partial extension of the classical weak sufficiency theory in the calculus of variations.

Key words. constrained minimization, function spaces, sufficient conditions, optimal control, nonnegative controls

AMS(MOS) subject classifications. 49K15, 46N10, 90C06

1. Introduction. Second-order sufficient conditions of the Kuhn–Tucker type are established here for constrained minimization problems

$$(1a) \quad \min_{D \cap \Omega^+} J,$$

where

$$(1b) \quad D \text{ is an open set in } \mathcal{L}^p(0, 1), \quad (p \in [2, \infty]),$$

$$(1c) \quad J : D \rightarrow \mathbb{R}^1,$$

$$(1d) \quad \Omega^+ = \{u \in \mathcal{L}^p(0, 1) : u(t) \geq 0 \text{ a.e.}\},$$

and J has first and second Gateaux differentials of the form

$$(2a) \quad d^1 J(u; v) = \int_0^1 \nabla J(u)(t) v(t) dt, \quad v \in \mathcal{L}^p(0, 1),$$

with

$$(2b) \quad \nabla J(u) \in \mathcal{L}^q(0, 1), \quad \left(\frac{1}{p} + \frac{1}{q} = 1 \right)$$

and

$$(2c) \quad d^2 J(u; v, w) = \int_0^1 [\nabla^2 J(u)v](t) w(t) dt, \quad v, w \in \mathcal{L}^p(0, 1),$$

with

$$(2d) \quad [\nabla^2 J(u)v](t) = S(u)(t)v(t) + \int_0^1 K(u)(t, s)v(s)ds, \quad t \in [0, 1],$$

* Received by the editors April 1, 1991; accepted for publication (in revised form) September 24, 1991. This research was supported by National Science Foundation grant DMS-9002848.

† Mathematics Department, Box 8205, North Carolina State University, Raleigh, North Carolina 27695-8205.

$$(2e) \quad S(u) \in \mathcal{L}^\infty(0, 1),$$

$$(2f) \quad K(u) \in \mathcal{L}^2([0, 1] \times [0, 1]),$$

and

$$(2g) \quad K(u)(t, s) = K(u)(s, t) \quad t, s \in [0, 1]$$

at each $u \in D$. In addition, we assume that

$$(3a) \quad \lim_{\substack{\|v-u\|_2 \rightarrow 0 \\ v \in D}} \|S(v) - S(u)\|_\infty = 0$$

and

$$(3b) \quad \lim_{\substack{\|v-u\|_2 \rightarrow 0 \\ v \in D}} \|K(v) - K(u)\|_2 = 0$$

at each $u \in D$; i.e., the maps $S : D \rightarrow \mathcal{L}^\infty(0, 1)$ and $K : D \rightarrow \mathcal{L}^2([0, 1] \times [0, 1])$ are continuous with respect to the \mathcal{L}^2 norm on $D \subset \mathcal{L}^p(0, 1)$ and the standard norms on $\mathcal{L}^\infty(0, 1)$ and $\mathcal{L}^2([0, 1] \times [0, 1])$. Under these circumstances, Taylor's theorem in \mathbb{R}^1 gives

$$(4) \quad J(u+v) = J(u) + d^1 J(u; v) + \frac{1}{2} d^2 J(u; v, v) + o(\|v\|_2^2), \quad v \in D$$

at each u in D . For $p = 2$, conditions (2) and (3) imply that J is twice continuously Fréchet differentiable on $D \subset \mathcal{L}^2(0, 1)$. On the other hand, for $p > 2$, the former conditions do not imply \mathcal{L}^2 differentiability on $D \subset \mathcal{L}^p(0, 1)$, since J need not be defined anywhere in $\mathcal{L}^2(0, 1) \sim D$.

The structure and smoothness assumptions in (2) and (3) are met in nontrivial classes of constrained-input optimal control problems with nonnegative admissible controls and Bolza objective functions

$$(5a) \quad J(u) = P(x(1)) + \int_0^1 f^o(t, x(t), u(t)) dt,$$

where $x(\cdot) : [0, 1] \rightarrow \mathbb{R}^n$ is the solution of an initial value problem

$$(5b) \quad x(0) = x_o$$

$$(5c) \quad \frac{dx}{dt} = f(t, x(t), u(t)), \quad t \in [0, 1].$$

When P , f^o , and f satisfy smoothness and growth restrictions suitably matched to $p \in [2, \infty]$, the initial value problem (5b), (5c) has a unique absolutely continuous solution $x(\cdot)$ corresponding to each $u(\cdot)$ in some open set $D \subset \mathcal{L}^p(0, 1)$, the associated Lebesgue integral in (5a) exists, and the composite function $J : D \rightarrow \mathbb{R}^1$ has Gateaux differentials (2a) and (2b) constructed as follows. Consider vectors in \mathbb{R}^m as $m \times 1$ column matrices, and for $(t, \psi, x, u) \in \mathbb{R}^1 \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^1$, put

$$(6) \quad H(t, \psi, x, u) = \psi^T f(t, x, u) + f^o(t, x, u),$$

where the superscript T denotes matrix transposition. Given $u \in D$ and the corresponding solution $x(\cdot)$ of (5b) and (5c), form the Jacobian matrices

$$(7a) \quad a(t) = \frac{\partial f^o}{\partial x}(t, x(t), u(t)),$$

$$(7b) \quad A(t) = \frac{\partial f}{\partial x}(t, x(t), u(t)),$$

$$(7c) \quad B(t) = \frac{\partial f}{\partial u}(t, x(t), u(t))$$

and let $\psi(\cdot) : [0, 1] \rightarrow \mathbb{R}^n$ denote the unique solution of the adjoint initial value problem

$$(8a) \quad \psi(1) = \nabla P(x(1)),$$

$$(8b) \quad \frac{d\psi}{dt} = -A(t)^T \psi - a(t)^T, \quad t \in [0, 1].$$

Compute the associated partial derivatives

$$(9) \quad \nabla J(u)(t) \triangleq \frac{\partial H}{\partial u}(t, \psi(t), x(t), u(t))$$

and Hessian matrices

$$(10a) \quad Q_1 = \nabla^2 P(x(1)),$$

$$(10b) \quad Q(t) = \nabla_{xx}^2 H(t, \psi(t), x(t), u(t)),$$

$$(10c) \quad R(t) = \nabla_{xu}^2 H(t, \psi(t), x(t), u(t)),$$

$$(10d) \quad S(t) = \frac{\partial^2 H}{\partial u^2}(t, \psi(t), x(t), u(t)).$$

Then, for v, w in $\mathcal{L}^p(0, 1)$,

$$(11a) \quad d^1 J(u; v) = \int_0^1 \nabla J(u)(t) v(t) dt$$

and

$$(11b) \quad \begin{aligned} d^2 J(u; v, w) = & z(1)^T Q_1 y(1) \\ & + \int_0^1 \{ z(t)^T Q(t) y(t) + z(t)^T R(t) v(t) \\ & + y(t)^T R(t) w(t) + S(t) w(t) v(t) \} dt, \end{aligned}$$

where $y(\cdot) : [0, 1] \rightarrow \mathbb{R}^n$ and $z(\cdot) : [0, 1] \rightarrow \mathbb{R}^n$ are the unique solutions of

$$(11c) \quad y(0) = 0,$$

$$(11d) \quad \frac{dy}{dt} = A(t)y + B(t)v(t), \quad t \in [0, 1],$$

$$(11e) \quad z(0) = 0,$$

$$(11f) \quad \frac{dz}{dt} = A(t)z + B(t)w(t), \quad t \in [0, 1]$$

(see [1]). To obtain (2c) and (2d) from (11b)–(11f), let $\Phi(\cdot, \tau)$ be the unique $n \times n$ matrix-valued solution of

$$(12a) \quad \Phi(\tau, \tau) = I,$$

$$(12b) \quad \frac{\partial \Phi}{\partial t}(t, \tau) = A(t)\Phi(t, \tau), \quad 0 \leq \tau \leq t \leq 1.$$

Then

$$(12c) \quad y(t) = \int_0^t \Phi(t, \tau)B(\tau)v(\tau)d\tau,$$

$$(12d) \quad z(t) = \int_0^t \Phi(t, \tau)B(\tau)w(\tau)d\tau,$$

and (11b)–(11f) yield (2c) and (2d), with $S(u)(t) = S(t)$ in (10d), and

$$(12e) \quad \begin{aligned} K(u)(t, s) &= B(t)^T \hat{\Phi}(s, t)^T R(s) + B(s)^T \hat{\Phi}(t, s)^T R(t) \\ &+ B(t)^T \left[\int_{\max(t, s)}^1 \Phi(\tau, t)^T Q(\tau) \Phi(\tau, s) d\tau \right] B(s) \\ &+ B(t)^T \Phi^T(1, t) Q_1 \Phi(1, s) B(s), \end{aligned}$$

where

$$(12f) \quad \hat{\Phi}(s, t) = \begin{cases} \Phi(s, t), & t \leq s, \\ 0, & s < t. \end{cases}$$

Once again, if P , f^o , and f are suitably restricted, then $\nabla J(u)$, $S(u)$, and $K(u)$ are defined at each $u \in D$ and possess the smoothness and symmetry properties specified in (2b), (2e)–(2g), and (3). We note that the related hypothesis of Fréchet differentiability for J on the normed vector space $\{\mathcal{L}^\infty(0, 1), \|\cdot\|_2\}$ has been invoked elsewhere in a control theoretic context [2].

From here onward, our analysis proceeds mainly at the level of (1)–(3). In §§2 and 3, we show that formal extensions of the Kuhn–Tucker second-order necessary

condition for local optimality in the nonnegative orthant in \mathbb{R}^n is also a necessary condition for \mathcal{L}^∞ -local optimality (and thus for \mathcal{L}^p -local optimality) in the set $D \cap \Omega^+$, but that a corresponding formal extension of the Kuhn–Tucker sufficient conditions in \mathbb{R}^n is generally not sufficient even for the weaker species of \mathcal{L}^∞ -local optimality in $D \cap \Omega^+$. These results amplify points already made in [3] for abstract nonlinear programs, $\min J(u)$ subject to $g(u) \geq 0$, where g has its range in some partially ordered infinite-dimensional Banach space; however, in §4, we prove a new result, namely, that the formal Kuhn–Tucker sufficient conditions do become sufficient for \mathcal{L}^∞ -local optimality at $u \in D \cap \Omega^+$ when the null set for u is closed in $[0, 1]$ and $S(u)$ is continuous on the frontier of this set in $[0, 1]$. In §2 we also establish a stronger variant of the Kuhn–Tucker second-order necessary condition for \mathcal{L}^p -local optimality in $D \cap \Omega^+$ with $p \in [2, \infty)$, and then prove in §4 that a natural further strengthening of this condition is sufficient for local optimality relative to the \mathcal{L}^2 norm on $D \cap \Omega^+$. Since these theorems have no counterparts in \mathbb{R}^n , they are a potential source of new insights for large-scale finite-dimensional approximations to (1)–(3), and, in particular, for multistage discrete-time approximations to (5) (see [4]–[6] for related illustrations of this point). In §5, we apply the sufficient conditions to three examples. Two of the examples are control problems; all three have nonconvex quadratic objective functions.

In the context of variational calculus, the results in §4 demonstrate that a significant portion of the classical weak sufficiency theory extends to certain input-constrained optimal control problems. In this same setting, we note that our results are not implied by the sufficient conditions in [7], nor are they contained in a recent extension of the Kuhn–Tucker sufficient conditions for nonlinear programs in the space $W(0, 1)$ of real functions with square-integrable derivatives [8]. Although portions of the sufficiency theory in [8] are applicable to constrained control and variational problems, the Kuhn–Tucker theorem proved there imposes a W -norm coercivity condition that cannot be satisfied when J has second differentials of the form (2c)–(2g).

As observed in [3] and elsewhere, convergence theories for iterative minimization algorithms often rest on sufficient conditions for local optimality, and, in particular, on second-order sufficient conditions of the Kuhn–Tucker type. In a sequel to the present article, it will be shown that the local convergence behavior of a standard gradient projection algorithm is indeed directly tied to the sufficient conditions in §4. These conclusions also have potentially interesting (computational) implications for finite-dimensional approximations to (1)–(3).

Finally, we hope and expect that the analysis in this paper serves as a model for a more general treatment of constrained minimization problems with objective functions J satisfying analogues of (2)–(3) on open sets in $\mathcal{L}^p([0, 1], \mathbb{R}^m)$, and with admissible function sets $\Omega = \mathcal{L}^p([0, 1], U)$, where U is polyhedral set in \mathbb{R}^m , or more generally, where U is prescribed by finitely many smooth inequality constraints in \mathbb{R}^m .

2. Necessary conditions. At each u in the set $D \cap \Omega^+$ put

$$(13a) \quad \alpha(u) = \{t \in [0, 1] : u(t) = 0\},$$

$$(13b) \quad T(u) = \{v \in \mathcal{L}^p(0, 1) : v(t) = 0 \text{ a.e. in } \alpha(u)\}.$$

The null sets $\alpha(u)$ are analogous to the active constraint index sets for the counterpart of (1) in \mathbb{R}^n , namely,

$$(14a) \quad \min_{D \cap \Omega_n^+} J,$$

$$(14b) \quad D \text{ is an open set in } \mathbb{R}^n,$$

$$(14c) \quad \Omega_n^+ = \{u \in \mathbb{R}^n : u_i \geq 0, i = 1, \dots, n\}.$$

Similarly, the closed subspaces $T(u)$ are analogous to the spaces tangent to the active constraint manifolds at u in Ω_n^+ . Note that if $u(t) = v(t)$ almost everywhere, then the symmetric difference $\alpha(u) \Delta \alpha(v) = [\alpha(u) \sim \alpha(v)] \cup [\alpha(v) \sim \alpha(u)]$ has measure zero, and $T(u) = T(v)$. The following theorems may now be seen as formal extensions of the Kuhn–Tucker first- and second-order necessary conditions for (14). We supply an elementary proof that is also suggested by the analogy with (14); however, results similar to those in Theorem 1 are known for a much larger class of nonlinear programs in an abstract Banach space setting [3].

THEOREM 1. *Let u be an \mathcal{L}^∞ -local minimizer of $J : D \rightarrow \mathbb{R}^1$ in the set $D \cap \Omega^+$; i.e.,*

$$(15) \quad u \in D \cap \Omega^+ \quad \text{and} \quad (\exists \rho > 0 \, \forall v \in D \cap \Omega^+, \|v - u\|_\infty < \rho \Rightarrow J(v) \geq J(u)).$$

In addition, suppose that the first and second Gateaux differentials of J exist at u , and that the associated maps $w \rightarrow d^1 J(u; w)$ and $w \rightarrow d^2 J(u; w, w)$ are continuous with respect to the \mathcal{L}^2 norm. Then

$$(16a) \quad \forall v \in \Omega^+ \quad d^1 J(u; v - u) \geq 0,$$

$$(16b) \quad \forall w \in T(u) \quad d^1 J(u; w) = 0,$$

$$(16c) \quad \forall w \in T(u) \quad d^2 J(u; w, w) \geq 0.$$

Proof. The set Ω^+ is convex, and D is open in $\mathcal{L}^p(0, 1)$; hence, if $u \in D \cap \Omega^+$ and $v \in \Omega^+$, then $u + \epsilon(v - u) \in D \cap \Omega^+$ for all sufficiently small $\epsilon > 0$. In view of (15), we therefore have

$$(17) \quad \forall v \in \Omega^+ \left(v - u \in \mathcal{L}^\infty(0, 1) \Rightarrow d^1 J(u; v - u) = \lim_{\epsilon \rightarrow 0^+} \frac{J(u + \epsilon(v - u)) - J(u)}{\epsilon} \geq 0 \right).$$

Condition (16a) now follows from (17) since $d^1 J(u; w)$ is \mathcal{L}^2 -continuous in w , and the set $\{v \in \Omega^+ : v - u \in \mathcal{L}^\infty\}$ is dense in Ω^+ .

To prove (16b) and (16c), we first define

$$\begin{aligned} \alpha_n(u) &= \left\{ t \in [0, 1] : 0 \leq u(t) < \frac{1}{n} \right\}, \\ T_n(u) &= \{v \in \mathcal{L}^p(0, 1) : v(t) = 0 \text{ a.e. in } \alpha_n(u)\}, \\ T_n^\infty(u) &= \{v \in \mathcal{L}^\infty(0, 1) : v(t) = 0 \text{ a.e. in } \alpha_n(u)\}, \end{aligned}$$

for $n = 1, 2, \dots$, and note that

$$\begin{aligned}\alpha_n(u) &\supset \alpha_{n+1}(u) \supset \alpha(u), \\ \bigcap_{n=1}^{\infty} [\alpha_n(u) \sim \alpha(u)] &= \phi, \\ \lim_{n \rightarrow \infty} \mu[\alpha_n(u) \sim \alpha(u)] &= 0 \quad (\mu = \text{Lebesgue measure,})\end{aligned}$$

and

$$T(u) \supset T_{n+1}(u) \supset T_n(u) \supset T_n^{\infty}(u),$$

with $T_n^{\infty}(u)$ dense in $T_n(u)$ and $\bigcup_{n=1}^{\infty} T_n(u)$ dense in $T(u)$. Suppose that $w \in T_n^{\infty}(u)$ with $w \neq 0$. By construction, $v = u \pm (n\|w\|_{\infty})^{-1}w \in \Omega^+$; hence (16a) gives

$$d^1 J(u; w) = n\|w\|_{\infty} d^1 J(u; v - u) \geq 0$$

and

$$d^1 J(u; w) = -n\|w\|_{\infty} d^1 J(u; v - u) \geq 0.$$

Since $d^1 J(u; 0) = 0$, we have shown that $d^1 J(u; w) = 0$ for all n , and all $w \in T_n^{\infty}(u)$. Condition (16b) now follows by continuous extension. Similarly, if $w \in T_n^{\infty}(u)$, then $v = u + \epsilon w \in D \cap \Omega^+$ for ϵ sufficiently small, in which case (15), (16b), and Taylor's formula give $0 \leq \frac{1}{2}\epsilon^2 d^2 J(u; w, w) + o(\epsilon^2)$. In the limit as $\epsilon \rightarrow 0$, we obtain $d^2 J(u; w, w) \geq 0$ for all n , and all $w \in T_n^{\infty}(u)$. Condition (16c) follows as before by continuous extension. \square

THEOREM 2. *Let u be an \mathcal{L}^{∞} -local minimizer of $J : D \rightarrow \mathbb{R}^1$ in the set $D \cap \Omega^+$, and suppose that J has first and second Gateaux differentials satisfying (2) at u . Then conditions (16) hold at u , and, consequently,*

$$(18a) \quad \nabla J(u)(t) \geq 0 \quad \text{a.e. in } \alpha(u),$$

$$(18b) \quad \nabla J(u)(t) = 0 \quad \text{a.e. in } \alpha(u)^c = [0, 1] \sim \alpha(u),$$

$$(18c) \quad S(u)(t) \geq 0 \quad \text{a.e. in } \alpha(u)^c.$$

Proof. Conditions (2) imply that $d^1 J(u; w)$ and $d^2 J(u; w, w)$ are \mathcal{L}^2 -continuous in w ; hence conditions (16) hold at u , by Theorem 1. By (16a), we have

$$\forall v \in \Omega^+ \quad \int_0^1 \nabla J(u)(t)[v(t) - u(t)]dt \geq 0$$

and, therefore,

$$(\forall v \geq 0, \nabla J(u)(t)[v - u(t)] \geq 0) \quad \text{a.e. in } [0, 1].$$

Assertions (18a) and (18b) now follow immediately from the definition of $\alpha(u)$.

To prove (18c), define $\theta(t, \epsilon) = \alpha(u)^c \cap (t - \epsilon, t + \epsilon)$ for $t \in \alpha(u)^c$ and $\epsilon > 0$. Since almost all points in $\alpha(u)^c$ are points of density for $\alpha(u)^c$ [9], and since $S(u) \in \mathcal{L}^\infty(0, 1)$, we have almost everywhere in $\alpha(u)^c$

$$\lim_{\epsilon \rightarrow 0^+} \frac{\mu(\theta(t, \epsilon))}{2\epsilon} = 1$$

and

$$\int_{\theta(t, \epsilon)} S(u)(\tau) d\tau = 2\epsilon S(u)(t) + o(\epsilon).$$

In addition, the Cauchy inequality gives

$$\left| \iint_{\theta(t, \epsilon) \times \theta(t, \epsilon)} K(u)(\tau, s) d\tau ds \right| \leq \left(\iint_{\theta(t, \epsilon) \times \theta(t, \epsilon)} K(u)(\tau, s)^2 d\tau ds \right)^{1/2} \mu(\theta(t, \epsilon)) = o(\epsilon)$$

for all $t \in \alpha(u)^c$. Since the characteristic function of $\theta(t, \epsilon)$ lies in the subspace $T(u)$, conditions (2) and (16c) now imply that

$$(2\epsilon S(u)(t) + o(\epsilon) \geq 0) \quad \text{a.e. in } \alpha(u)^c,$$

and this yields (18c) in the limit as $\epsilon \rightarrow 0^+$. \square

Note 1. Theorems 1 and 2 remain valid if u is merely an internal point [10] of the (otherwise arbitrary) set D and if (15) is replaced by the weaker requirement of local optimality in $(u + \text{span}\{w\}) \cap D \cap \Omega^+$ for all $w \in \mathcal{L}^\infty$.

Note 2. The counterpart of Theorem 2 for problem (14) asserts that, if u is a local minimizer of J in $D \cap \Omega^+$ (relative to any norm in \mathbb{R}^n) then,

$$(19a) \quad \forall v \in \Omega^+, \quad \sum_{i=1}^n \frac{\partial J}{\partial u_i}(u)(v_i - u_i) \geq 0$$

$$(19b) \quad \forall i \in \alpha(u) = \{i \in \{1, \dots, n\} : u_i = 0\}, \quad \frac{\partial J}{\partial u_i}(u) \geq 0$$

$$(19c) \quad \forall i \in \alpha(u)^c = \{1, \dots, n\} \sim \alpha(u), \quad \frac{\partial J}{\partial u_i}(u) = 0$$

$$(19d) \quad \forall w \in T(u) = \{w \in \mathbb{R}^n : \forall i \in \alpha(u), w_i = 0\}, \quad \sum_{i=1}^n \sum_{j=1}^n w_i \frac{\partial^2 J}{\partial u_i \partial u_j}(u) w_j \geq 0$$

$$(19e) \quad \forall i \in \alpha(u)^c, \quad \frac{\partial^2 J}{\partial u_i \partial u_i}(u) \geq 0.$$

From this vantage point, we see that $S(u)$ acts like the “diagonal part” of the Hessian operator $\nabla^2 J(u)$ in (2).

Note 3. For input-constrained optimal control problems (1) with objective functions (5), Theorem 2 yields the following extension of the Legendre–Clebsch condition in the variational calculus and unconstrained optimal control theory [11], [12]

$$(20) \quad \frac{\partial^2 H}{\partial u^2}(t, \psi(t), x(t), u(t)) \geq 0 \quad \text{a.e. in } \alpha(u)^c$$

(cf. (6)–(10)). This result and conditions (18a) and (18b) are also implied by the Pontryagin minimum principle [11], [12],

$$(21) \quad H(t, \psi(t), x(t), u(t)) = \min_{v \geq 0} H(t, \psi(t), x(t), v) \quad \text{a.e. in } [0, 1]$$

when $p = \infty$.

For $p \in [2, \infty]$, the first- and second-order necessary conditions in Theorem 2 automatically hold at any \mathcal{L}^p -local minimizer of J in $D \cap \Omega^+$; moreover, for $p \in [2, \infty)$, we can prove a stronger version of the second order condition (18c).

THEOREM 3. *Let the hypotheses of Theorem 2 hold with $p \in [2, \infty)$, and suppose that u is an \mathcal{L}^p -local minimizer of J in Ω^+ ; i.e.,*

$$(22) \quad u \in D \cap \Omega^+ \quad \text{and} \quad (\exists \rho > 0 \, \forall v \in D \cap \Omega^+, \|v - u\|_p < \rho \Rightarrow J(v) \geq J(u)).$$

Then conditions (16) and (18) are satisfied at u , and, in addition,

$$(23) \quad S(u)(t) \geq 0 \quad \text{a.e. in } [0, 1].$$

Proof. Every \mathcal{L}^p -local minimizer is also an \mathcal{L}^∞ -local minimizer; hence (16) and (18) hold at u , by Theorem 2. To prove (23), we must therefore show that

$$(24) \quad S(u)(t) \geq 0 \quad \text{a.e. in } \alpha(u).$$

As in the proof of Theorem 2, we construct $\theta(t, \epsilon) = \alpha(u) \cap (t - \epsilon, t + \epsilon)$ for $t \in \alpha(u)$ and $\epsilon > 0$, and find that

$$\left(\int_{\theta(t, \epsilon)} S(u)(\tau) d\tau = 2\epsilon S(t) + o(\epsilon) \right) \quad \text{a.e. in } \alpha(u)$$

and

$$\left| \iint_{\theta(t, \epsilon) \times \theta(t, \epsilon)} K(\tau, s) d\tau ds \right| = o(\epsilon)$$

for all $t \in \alpha(u)$. Furthermore, in view of (18b), we have

$$\left(\int_{\theta(t, \epsilon)} \nabla J(u)(\tau) d\tau = 2\epsilon \nabla J(u)(t) + o(\epsilon) \right) \quad \text{a.e. in } \alpha(u).$$

For $h > 0$ and $\epsilon > 0$, construct $v_{h, \epsilon} \in \Omega^+$ by the rule

$$v_{h, \epsilon}(\tau) = \begin{cases} h, & \tau \in \theta(t, \epsilon), \\ 0, & \tau \in \theta(t, \epsilon)^c. \end{cases}$$

Then

$$\forall h > 0 \quad \|v_{h, \epsilon}\|_p = O(\epsilon^{1/p})$$

and

$$\forall h \exists \epsilon_h \forall \epsilon \in (0, \epsilon_h) \quad u + v_{h, \epsilon} \in D \cap \Omega^+ \quad \text{and} \quad J(u + v_{h, \epsilon}) - J(u) \geq 0.$$

Conditions (2), (4), and the preceding estimates then give

$$(\forall h > 0, \quad 0 \leq [2\nabla J(u)(t) + hS(u)(t)]\epsilon + o(\epsilon)) \quad \text{a.e. in } \alpha(u).$$

In the limit as $\epsilon \rightarrow 0^+$, we therefore have

$$(\forall h > 0, \quad 0 \leq [2\nabla J(u)(t) + hS(u)(t)]) \quad \text{a.e. in } \alpha(u),$$

and, in the limit as $h \rightarrow \infty$, we obtain (24). \square

Note 4. The finite-dimensional counterpart of condition (23) is not necessary for local optimality in problem (14).

Note 5. Condition (23) is implied by the Pontryagin minimum principle (21) for optimal control problems (1) with objective functions (5).

3. Formal Kuhn–Tucker sufficient conditions. The following stronger variants of (18a), (18b), and (16c) amount to a formal extension of the standard Kuhn–Tucker sufficient conditions for (14):

$$(25a) \quad \nabla J(u)(t) \geq 0 \quad \text{a.e. in } \alpha(u),$$

$$(25b) \quad \forall \beta \subset \text{INT } \alpha(u) \quad (\beta \text{ compact} \Rightarrow \exists c_1 > 0, \nabla J(u)(t) \geq c_1 \quad \text{a.e. in } \beta),$$

$$(25c) \quad \nabla J(u)(t) = 0 \quad \text{a.e. in } \alpha(u)^c,$$

$$(25d) \quad \exists c > 0 \quad \forall v \in T(u), \quad d^2 J(u; v, v) \geq c \|v\|_2^2.$$

However, in the absence of further restrictions on the differential $d^2 J$, these conditions are actually not sufficient for \mathcal{L}^∞ -local optimality in the set $D \cap \Omega^+$, even when (2) and (3) hold.

Example 1. For $u \in D = \mathcal{L}^2(0, 1)$, put

$$(26) \quad J(u) = \int_0^1 \left[r(t)u(t) + \frac{1}{2}S(t)u(t)^2 \right] dt,$$

where $r(\cdot)$ is continuous and strictly decreasing with $r(\frac{1}{2}) = 0$ and

$$S(t) = \begin{cases} -1, & t \in [0, \frac{1}{2}), \\ 1, & t \in [\frac{1}{2}, 1]. \end{cases}$$

Then, for $v \in \mathcal{L}^2(0, 1)$,

$$d^1 J(u; v) = \int_0^1 [r(t) + S(t)u(t)] v(t) dt$$

and

$$d^2 J(u; v, v) = \int_0^1 S(t)v^2(t) dt.$$

In particular, for

$$(27) \quad u(t) = \begin{cases} 0, & t \in [0, \frac{1}{2}), \\ -r(t), & t \in (\frac{1}{2}, 1], \end{cases}$$

we have

$$u(\cdot) \in D \cap \Omega^+ = \Omega^+,$$

$$\alpha(u) = [0, \tfrac{1}{2}],$$

$$T(u) = \{v \in \mathcal{L}^2(0, 1) : v(t) = 0 \text{ a.e. in } [0, \tfrac{1}{2}]\},$$

$$\nabla J(u)(t) = \begin{cases} r(t), & t \in [0, \tfrac{1}{2}], \\ 0, & t \in (\tfrac{1}{2}, 1], \end{cases}$$

and

$$\forall v \in T(u) \quad d^2 J(u; v, v) = \int_{1/2}^1 v^2(t) dt = \|v\|_2^2.$$

Consequently, conditions (2), (3), and (25) hold at $u(\cdot)$. According to (23), we already know that J cannot have \mathcal{L}^2 -local minimizers in Ω^+ ; however, $u(\cdot)$ is also not an \mathcal{L}^∞ -local minimizer for J in Ω^+ . To see this, let

$$t_n = \frac{1}{2} - \frac{1}{n}$$

for $n = 1, 2, \dots$, and put

$$v_n(t) = \begin{cases} 2r(t_n), & t \in [t_n, \tfrac{1}{2}), \\ 0, & t \in [t_n, \tfrac{1}{2})^c. \end{cases}$$

By construction,

$$J(u + v_n) - J(u) = 2r(t_n) \int_{t_n}^{1/2} [r(t) - r(t_n)] dt < 0,$$

$$u + v_n \in \Omega,$$

and

$$\lim_{n \rightarrow \infty} \|v_n\|_\infty = 2 \lim_{n \rightarrow \infty} r(t_n) = 0.$$

Hence u is not an \mathcal{L}^∞ -local minimizer.

Example 1 demonstrates two basic points: First, u cannot be \mathcal{L}^∞ -locally optimal for the simplest quadratic J satisfying (2) and (3) if there is a t^* in $\alpha(u)$ such that $\nabla J(u)(t)$ approaches zero as t approaches t^* within $\alpha(u)$, while $S(u)(t) = S(t)$ remains negative and bounded away from zero in $\alpha(u)$ near t^* . Second, conditions (25) alone cannot eliminate such t^* . On the other hand, the next example suggests that (25) may become sufficient for \mathcal{L}^∞ -local optimality when $S(u)$ is continuous on the frontier of $\alpha(u)$ in $[0, 1]$.

Example 2. For $v \in D = \mathcal{L}^2(0, 1)$ define J by (26), with r as before, but

$$S(t) = \begin{cases} -1, & t \in [0, \tfrac{1}{2} - \delta], \\ 1, & t \in (\tfrac{1}{2} - \delta, 1], \end{cases}$$

for δ fixed in $(0, \frac{1}{2})$. Construct $u \in \Omega^+$ as in (27). Then conditions (25) hold once again at u , but now $S(u)(t) = +1$ on the open neighborhood $(\frac{1}{2} - \delta, 1]$ of $\alpha(u)^c = (\frac{1}{2}, 1]$, and this is enough to ensure that u is an \mathcal{L}^∞ -local minimizer. More specifically, let $\rho = r(\frac{1}{2} - \delta) > 0$, and suppose that $u + v \in \Omega^+$ and $\|v\|_\infty < \rho$. Then

$$\begin{aligned} J(u+v) - J(u) &= \int_0^1 r(t)v(t)dt + \frac{1}{2} \int_0^1 S(t) (v(t)^2 + 2u(t)v(t)) dt \\ &= \int_0^{\frac{1}{2}-\delta} r(t)v(t)dt + \int_{\frac{1}{2}-\delta}^{\frac{1}{2}} r(t)v(t)dt \\ &\quad - \frac{1}{2} \int_0^{\frac{1}{2}-\delta} v(t)^2 dt + \frac{1}{2} \int_{\frac{1}{2}-\delta}^1 v(t)^2 dt \\ &\geq \int_0^{\frac{1}{2}-\delta} \left[\rho - \frac{1}{2}v(t) \right] v(t)dt + \frac{1}{2} \int_{\frac{1}{2}-\delta}^1 v(t)^2 dt \\ &\geq \frac{1}{2} \int_0^1 v(t)^2 dt \\ &= \frac{1}{2} \|v\|_2^2. \end{aligned}$$

Note that u still cannot be an \mathcal{L}^p -local minimizer of J in Ω^+ for $p \in [2, \infty)$ since condition (23) is still violated at u (however, if we set $S(t) = 1$ in $[0, 1]$ then (23) is satisfied and (27) does, in fact, define the unique global minimizer of the now strictly convex functional J in Ω^+).

In Example 2, the crucial fact is that $S(u)$ and $\nabla J(u)$ are positive and bounded away from zero on some open neighborhood \mathcal{O}_δ of $\alpha(u)^c$ and on \mathcal{O}_δ^c , respectively. For simple quadratic functionals (26) with $K = 0$, these conditions can be inferred from (25) and continuity restrictions on $S(u)$; moreover, with a straightforward extension of the estimates in Example 2, we can show that this further strengthening of (25) is sufficient for \mathcal{L}^∞ -local optimality in $D \cap \Omega^+$. In the next section, we establish analogous general results for the class of nonquadratic J satisfying (2) and (3) with $K \neq 0$, and we also prove that the sufficient conditions for \mathcal{L}^∞ -local optimality become sufficient conditions for \mathcal{L}^2 -local optimality when $K(u) \in \mathcal{L}^\infty([0, 1] \times [0, 1])$ and $S(u)(t)$ is bounded away from zero almost everywhere in $[0, 1]$ (cf. Theorem 3).

4. Sufficient conditions. We begin with several definitions and a fundamental sufficiency lemma for (1).

For measurable $\alpha \subset [0, 1]$, put

$$T_\alpha = \{w \in \mathcal{L}^p(0, 1) : w(t) = 0 \text{ a.e. in } \alpha\},$$

$$N_\alpha = \{w \in \mathcal{L}^p(0, 1) : w(t) = 0 \text{ a.e. in } \alpha^c\}.$$

For $v \in \mathcal{L}^p(0, 1)$ define the corresponding projections of v into T_α and N_α by

$$\begin{aligned} P_{T_\alpha} v &= \begin{cases} 0, & t \in \alpha, \\ v(t), & t \in \alpha^c, \end{cases} \\ P_{N_\alpha} v &= \begin{cases} v(t), & t \in \alpha, \\ 0, & t \in \alpha^c. \end{cases} \end{aligned}$$

By construction, $v = P_{T_\alpha} v + P_{N_\alpha} v$ and

$$P_{T_\alpha} v(t) P_{N_\alpha} v(t) dt = 0 \quad t \in [0, 1].$$

LEMMA 1. *Let the first and second Gateaux differentials of $J : D \rightarrow \mathbb{R}^1$ exist at $u \in D \cap \Omega^+$ and satisfy (4), with $d^1 J(u; v)$ linear in v , and $d^2 J(u; v, w)$ bilinear in (v, w) and bounded relative to the \mathcal{L}^2 -norm; i.e.,*

$$(28) \quad \exists M \geq 0 \quad \forall (v, w) \in \mathcal{L}^p(0, 1) \times \mathcal{L}^p(0, 1) \quad |d^2 J(u; v, w)| \leq M \|v\|_2 \|w\|_2.$$

Suppose that u satisfies the necessary conditions (16), and, moreover, that there are positive numbers c_1 and c_2 , and a measurable set α such that

$$(29a) \quad \mu[\alpha \sim \alpha(u)] = 0,$$

$$(29b) \quad \forall w \in N_\alpha \cap \Omega^+, \quad d^1 J(u; w) \geq c_1 \|w\|_1,$$

$$(29c) \quad \forall w \in T_\alpha \supset T(u), \quad d^2 J(u; w, w) \geq c_2 \|w\|_2^2.$$

Then u is a strict \mathcal{L}^∞ -local minimizer for J in $D \cap \Omega^+$; more specifically,

$$(30) \quad \exists \rho > 0 \quad \exists d > 0 \quad \forall v \in D \cap \Omega^+ \quad (\|v - u\|_\infty < \rho \Rightarrow J(v) - J(u) \geq d \|v - u\|_2^2).$$

Proof. For $u + w \in D \cap \Omega^+$, conditions (4), (28), and (29) give

$$\begin{aligned} J(u+w) - J(u) &= d^1 J(u; P_{T_\alpha} w + P_{N_\alpha} w) + \frac{1}{2} d^2 J(u; P_{T_\alpha} w + P_{N_\alpha} w, P_{T_\alpha} w + P_{N_\alpha} w) \\ &\quad + o(\|v\|_2^2) \\ &\geq c_1 \|P_{N_\alpha} w\|_1 + d^1 J(u; P_{T_\alpha} w) + \frac{1}{2} c_2 \|P_{T_\alpha} w\|_2^2 \\ &\quad - M \left(\frac{1}{2} \|P_{N_\alpha} w\|_2 + \|P_{T_\alpha} w\|_2 \right) \|P_{N_\alpha} w\|_2 + o(\|w\|_2^2). \end{aligned}$$

By (29a), we also have, for all $w \in \mathcal{L}^p(0, 1)$, that

$$P_{T_\alpha} w = P_{T_{\alpha(u)}} w + P_{T_{[\alpha(u) \sim \alpha]^c}} w \quad \text{a.e. in } [0, 1],$$

with $P_{T_{\alpha(u)}} w \in T_{\alpha(u)} = T(u)$ and

$$u + w \in \Omega^+ \Rightarrow (u + P_{T_{[\alpha(u) \sim \alpha]^c}} w) \in \Omega^+.$$

Consequently, if $u + w \in \Omega^+$, then (16a) and (16b) yield

$$d^1 J(u; P_{T_\alpha} w) = d^1 J(u; P_{T_{\alpha(u)}} w) + d^1 J(u; P_{T_{[\alpha(u) \sim \alpha]^c}} w) \geq 0.$$

Thus, for all $w \in \mathcal{L}^\infty(0, 1)$,

$$\begin{aligned}
& u + w \in D \cap \Omega^+ \quad \text{and} \quad \|w\|_\infty < \rho \Rightarrow \\
& J(u + w) - J(u) \geq c_1 \rho^{-1} \|P_{N_\alpha} w\|_2^2 + \frac{1}{2} c_2 \|P_{T_\alpha} w\|_2^2 \\
& \quad - M \left(\frac{1}{2} \|P_{N_\alpha} w\|_2 + \|P_{T_\alpha} w\|_2 \right) \|P_{N_\alpha} w\|_2 \\
& \quad + o(\|w\|_2^2) \\
(31) \quad & = \left(c_1 \rho^{-1} - \frac{1}{2} M \right) \|P_{N_\alpha} w\|_2^2 + \frac{1}{4} c_2 \|P_{T_\alpha} w\|_2^2 \\
& \quad + \frac{1}{4} c_2 \left(\|P_{T_\alpha} w\|_2^2 - 4 M c_2^{-1} \|P_{N_\alpha} w\|_2 \|P_{T_\alpha} w\|_2 \right) \\
& \quad + o(\|w\|_2^2) \\
& \geq \left(c_1 \rho^{-1} - \frac{1}{2} M - M_2 c_2^{-1} \right) \|P_{N_\alpha} w\|_2^2 + \frac{1}{4} c_2 \|P_{T_\alpha} w\|_2^2 \\
& \quad + o(\|w\|_2^2).
\end{aligned}$$

If we now choose $\rho > 0$ so small that

$$(32a) \quad c_1 \rho^{-1} - \frac{1}{2} M - M_2 c_2^{-1} \geq \frac{1}{4} c_2,$$

then

$$(32b) \quad J(u + w) - J(u) \geq \frac{1}{4} c_2 \|w\|_2^2 + o(\|w\|_2^2)$$

for all $w \in \mathcal{L}^\infty(0, 1)$ such that $u + w \in D \cap \Omega^+$ and $\|w\|_\infty < \rho$. Estimate (30) follows at once from (32). \square

Our objective now is to establish sufficient conditions of the Kuhn–Tucker type for (1)–(3) by deducing (29) from the formal sufficient conditions (25) and additional continuity restrictions on $S(u)$ in (2). Along with Lemma 1, we need the following results.

LEMMA 2. *Suppose that $J : D \rightarrow \mathbb{R}^1$ has a second Gateaux differential satisfying (2c)–(2g) at $u \in D$, and let (25d) hold at u ; i.e.,*

$$\exists c > 0 \quad \forall v \in T(u) \quad d^2 J(u; v, v) \geq c \|v\|_2^2.$$

Then

$$(33) \quad S(u)(t) \geq c \quad \text{a.e. in } \alpha(u)^c.$$

Moreover, if

$$(34) \quad \mu \left(\overline{\alpha(u)^c} \sim \alpha(u)^c \right) = 0,$$

and if (33) extends to some open set \mathcal{O}_o in $[0, 1]$ containing $\overline{\alpha(u)^c}$, then there is another open set \mathcal{O} in $[0, 1]$ such that

$$(35a) \quad \mathcal{O}_o \supset \mathcal{O} \supset \overline{\alpha(u)^c}$$

and

$$(35b) \quad \forall v \in T_\alpha \supset T(u) \quad d^2 J(u; v, v) \geq \frac{1}{2} c \|v\|_2^2,$$

with

$$(35c) \quad \alpha = \mathcal{O}^c.$$

Proof. Condition (33) is established by a trivial modification of the proof for (18c). Now suppose that the stronger condition

$$(36) \quad S(u)(t) \geq c \quad \text{a.e. in } \mathcal{O}_o \supset \overline{\alpha(u)^c}$$

actually holds. Note that, for $\mathcal{O}_o \supset \mathcal{O} \supset \overline{(\alpha(u))^c}$, $\alpha = \mathcal{O}^c$ and $v \in T_\alpha$, we have

$$\begin{aligned} d^2 J(u; v, v) &= \int_{\mathcal{O}} S(u)(t) v(t)^2 dt + \iint_{\mathcal{O} \times \mathcal{O}} K(u)(t, s) v(t) v(s) dt ds \\ &= \int_{\overline{\alpha(u)^c}} S(u)(t) v(t)^2 dt + \iint_{\overline{\alpha(u)^c} \times \overline{\alpha(u)^c}} K(u)(t, s) v(t) v(s) dt ds \\ &\quad + \int_{\mathcal{O} \sim \overline{\alpha(u)^c}} S(u)(t) v(t)^2 dt + \iint_{\mathcal{O} \times \mathcal{O}} \widehat{K}(u)(t, s) v(t) v(s) dt ds \end{aligned}$$

with

$$\widehat{K}(u)(t, s) = \begin{cases} 0, & (t, s) \in \overline{\alpha(u)^c} \times \overline{\alpha(u)^c}, \\ K(u)(t, s), & (t, s) \in \Delta \mathcal{O} \end{cases}$$

and

$$\Delta \mathcal{O} = (\mathcal{O} \times \mathcal{O}) \sim \left(\overline{\alpha(u)^c} \times \overline{\alpha(u)^c} \right).$$

Furthermore, by (34), $\mu \left[\left(\overline{\alpha(u)^c} \times \overline{\alpha(u)^c} \right) \sim (\alpha(u)^c \times \alpha(u)^c) \right] = 0$. Finally, since $\overline{\alpha(u)^c}$ is measurable, there is a sequence of open sets \mathcal{O}_n such that $\mathcal{O}_o \supset \mathcal{O}_n \supset \overline{\alpha(u)^c}$ and $\mu(\Delta \mathcal{O}_n) \rightarrow 0$. Hence there is an open set \mathcal{O} such that $\mathcal{O}_o \supset \mathcal{O} \supset \overline{\alpha(u)^c}$ and

$$\iint_{\mathcal{O} \times \mathcal{O}} \widehat{K}(u)(t, s)^2 dt ds = \iint_{\Delta \mathcal{O}} K(u)(t, s)^2 dt ds \leq \left(\frac{c}{2} \right)^2.$$

For any such \mathcal{O} , conditions (25d), (36), Cauchy's inequality, and the preceding estimates yield

$$\begin{aligned} d^2 J(u; v, v) &\geq c \int_{\overline{\alpha(u)^c}} v(t)^2 dt + c \int_{\mathcal{O} \sim \overline{\alpha(u)^c}} v(t)^2 dt - \frac{c}{2} \int_{\mathcal{O}} v(t)^2 dt \\ &= \frac{c}{2} \|v\|_2^2 \end{aligned}$$

for all $v \in T_\alpha \supset T(u)$, with $\alpha = \mathcal{O}^c$. \square

LEMMA 3. Suppose that β is an open set in $[0, 1]$ and that $f : [0, 1] \rightarrow \mathbb{R}^1$ is continuous on the frontier of β in $[0, 1]$. Moreover, suppose that, for some $c > 0$,

$$(37) \quad f(t) \geq c \quad \text{a.e. in } \beta.$$

Then there is an open set \mathcal{O}_o in $[0, 1]$ such that

$$(38a) \quad \mathcal{O}_o \supset \bar{\beta}$$

and

$$(38b) \quad f(t) \geq \tfrac{1}{2}c \quad \text{a.e. in } \mathcal{O}_o.$$

Proof. Note that $\beta = \text{INT } \beta$ and $\partial\beta = \bar{\beta} \sim \text{INT } \beta$. If $\partial\beta = \emptyset$, then $\bar{\beta} = \text{INT } \beta = [0, 1]$, and (38) follows trivially from (37) with $\mathcal{O} = [0, 1]$. On the other hand, suppose that $t \in \partial\beta$. Since β is open in $[0, 1]$, we have, for sufficiently small $\epsilon < 0$,

$$\mu[\beta \cap (t - \epsilon, t + \epsilon)] > 0.$$

Therefore, by (37), there is a sequence of points $t_n \in \beta$ such that $t_n \rightarrow t$ and $f(t_n) \geq c$ for all n . Since f is continuous at t , we must have $f(t) \geq c$, and therefore

$$\exists \delta_t > 0 \quad \forall \tau \in B(t, \delta_t) \quad f(\tau) \geq \tfrac{1}{2}c$$

with

$$B(t, \delta_t) = \{\tau \in [0, 1] : |\tau - t| < \delta_t\}.$$

It follows that (38) holds in the open set $\mathcal{O}_o = \beta \cup [\cup_{t \in \partial\beta} B(t, \delta_t)] \supset \beta \cup \partial\beta = \bar{\beta}$. \square

With Lemmas 1–3 established, we can now prove the following extension of the Kuhn–Tucker sufficient conditions.

THEOREM 4. Let $J : D \rightarrow \mathbb{R}^1$ have first and second Gateaux differentials satisfying (2) and (3) at $u \in D \cap \Omega^+$, and suppose that the formal Kuhn–Tucker conditions (25) are satisfied at u . In addition, assume that $\alpha(u)$ is closed in $[0, 1]$ and that $S(u)$ is continuous on the frontier of $\alpha(u)$ in $[0, 1]$. Then $S(u)(t)$ is bounded away from zero almost everywhere in some open neighborhood of $\overline{\alpha(u)^c}$ in $[0, 1]$, u is a strict \mathcal{L}^∞ -local minimizer of J in $D \cap \Omega^+$, and condition (30) is satisfied; moreover, if $K(u) \in \mathcal{L}^\infty([0, 1] \times [0, 1])$ and $S(u)$ is positive and bounded away from zero almost everywhere in $[0, 1]$, then u is also strictly locally optimal relative to the \mathcal{L}^2 norm on $D \cap \Omega^+$; more specifically, $u \in D \cap \Omega^+$ and

$$(39) \quad \exists \rho > 0 \quad \exists d > 0 \quad \forall v \in D \cap \Omega^+ \quad (\|v - u\|_2 < \rho \Rightarrow J(v) - J(u) \geq d\|v - u\|^2).$$

Proof. We show that the hypotheses in Lemma 1 are satisfied. By (2) and (3), $d^1 J(u; v)$ is linear in $v \in \mathcal{L}^p(0, 1)$, $d^2 J(u; v, w)$ is bilinear in $(v, w) \in \mathcal{L}^p(0, 1) \times \mathcal{L}^p(0, 1)$ and satisfies (28), and condition (4) holds. In view of Lemma 2, conditions (25) imply (33), as well as (16). The set $\alpha(u)^c$ is open, and hence (34) holds. Since $S(u)$ is also continuous on $\partial[\alpha(u)^c]$ and Lemma 3 establishes the existence of an open set \mathcal{O}_o in $[0, 1]$ such that $\mathcal{O}_o \supset \overline{\alpha(u)^c}$ and $S(u)(t) \geq \tfrac{1}{2}c$ almost everywhere in \mathcal{O}_o . Consequently, by the second part of Lemma 2, there is an open set \mathcal{O} in $[0, 1]$ such that $\mathcal{O}_o \supset \mathcal{O} \supset \overline{\alpha(u)^c}$ and conditions (29a) and (29c) hold with $\alpha = \mathcal{O}^c$ and $c_2 = \tfrac{1}{4}c$.

In addition, \mathcal{O}^c is a compact subset of $(\overline{\alpha(u)^c})^c = \text{INT } \alpha(u)$; hence (2a), (2b), and (25b) imply that (29b) also holds at u with $\alpha = \mathcal{O}^c$. Therefore, by Lemma 1, u is an \mathcal{L}^∞ -local minimizer of J in $D \cap \Omega^+$, and condition (30) is satisfied. Furthermore, if $K(u)$ is essentially bounded and $S(u)(t)$ is positive and bounded below by $c_3 > 0$ almost everywhere in $[0, 1]$, then, with reference to (2)–(4), (25), and the proof of Lemma 1, we find that, for all w such that $u + w \in D \cap \Omega^+$,

$$\begin{aligned} J(u+w) - J(u) &\geq d^1 J(u; P_{N_\alpha} w) + \frac{1}{2} \int_0^1 S(u)(t) [P_{N_\alpha} w(t) + P_{T_\alpha} w(t)]^2 dt \\ &\quad + \frac{1}{2} \int_0^1 \int_0^1 K(u)(t, s) [P_{N_\alpha} w(t) + P_{T_\alpha} w(t)] [P_{N_\alpha} w(s) + P_{T_\alpha} w(s)] dt ds \\ &\quad + o(\|w\|_2^2) \\ &\geq c_1 \|P_{N_\alpha} w\|_1 + \frac{1}{2} c_3 \|P_{N_\alpha} w\|_2^2 + \frac{1}{2} c_2 \|P_{T_\alpha} w\|_2^2 \\ &\quad - \|K(u)\|_\infty \left(\|P_{T_\alpha} w\|_1 + \frac{1}{2} \|P_{N_\alpha} w\|_1 \right) (\|P_{N_\alpha} w\|_1) + o(\|w\|_2^2) \\ &\geq \left(c_1 - \frac{3}{2} \|K(u)\|_\infty \|w\|_2 \right) \|P_{N_\alpha} w\|_1 + 2d \|w\|_2^2 + o(\|w\|_2^2), \end{aligned}$$

with $\alpha = \mathcal{O}^c$ and $d = \frac{1}{4} \min\{c_2, c_3\}$. Consequently, condition (39) is satisfied with some positive $\rho \leq \frac{2}{3} c_1 \|K(u)\|_\infty^{-1}$. \square

Note 6. The set $\alpha(u)$ is closed if u is lower semicontinuous. Conditions (25a) and (25b) hold if $\alpha(u)$ is closed and $\nabla J(u)$ is positive and lower semicontinuous on $\text{INT } \alpha(u)$.

Note 7. The topological restrictions on $\alpha(u)$ and $S(u)$ can be relaxed in Theorem 4 if we are willing to *assume* that $S(u)$ is bounded away from zero on some open neighborhood of $\overline{\alpha(u)^c}$ in $[0, 1]$; however, the resulting \mathcal{L}^∞ -local optimality sufficient conditions are then further removed from the necessary conditions in Theorem 2.

Note 8. In the variational calculus and unconstrained optimal control theory, \mathcal{L}^2 coercivity conditions on the second variation (i.e., $d^2 J(u; v, v)$) are deduced from the strict Legendre-Clebsch condition $((\partial^2 H / \partial u^2)(t, \psi(t), x(t), u(t)) = S(u)(t) > 0)$ and the Jacobi conjugate point condition [11], [12]. We may therefore ask whether (24d) can be inferred from (33) and from some natural generalization of the classical field embedding and Jacobi's condition for constrained input optimal control problems with Bolza objective functions and nonnegative admissible controls. A theorem of this kind and the \mathcal{L}^∞ -local optimality results in this section would complete the extension of the variational weak sufficiency theory to (1)–(3) and (5).

Note 9. When J has Gateaux differentials of the form (2), it is not possible to base a counterpart of the sufficiency theory constructed here on weaker versions of the continuity hypothesis (3) yielding

$$\exists p \in (2, \infty) \quad J(u+v) = J(u) + d^1 J(u; v) + \frac{1}{2} d^2 J(u; v, v) + o(\|v\|_p^2)$$

in place of (4). Such a theory would require stronger \mathcal{L}^p analogues of the coercivity condition (25d) that cannot hold for (2c)–(2g) since $|d^2 J(u; v, v)| \leq M \|v\|_2^2$ for some $M \geq 0$, while $\sup_{\|v\|_2=1} \|v\|_p = \infty$ for all $p > 2$. Similarly, in the present setting

the $W(0, 1)$ Kuhn–Tucker sufficient condition in [8] requires the coercivity condition (25d), with $\|v\|_2^2$ replaced by

$$\|v\|_W^2 = |v(0)|^2 + \int_0^1 \left| \frac{dv}{dt}(t) \right|^2 dt,$$

and this is also incompatible with (2c)–(2g), since $\sup_{\|v\|_2=1} \|v\|_W = \infty$.

5. Examples. In this section, we apply the results developed in previous sections to three optimization examples that illustrate various features of the theory. The first two examples are optimal control problems in the form

$$(40a) \quad \min J(u) = \int_0^1 [r(t)u(t) + \frac{1}{2}Q(t)x(t)^2 + \frac{1}{2}S(t)u(t)^2] dt,$$

$$(40b) \quad \dot{x}(t) = A(t)x(t) + B(t)u(t), \quad x(0) = 0,$$

$$(40c) \quad u \in \Omega^+ = \{w \in \mathcal{L}^2(0, 1) : w(t) \geq 0 \text{ a.e. in } [0, 1]\},$$

with a stationary control

$$(40d) \quad u^*(t) = \begin{cases} 0, & t \in [0, \frac{1}{2}], \\ 2t - 1, & t \in (\frac{1}{2}, 1]. \end{cases}$$

In our first example, we choose r , S , Q , A , and B , so that u^* is an \mathcal{L}^∞ -local minimizer, but not an \mathcal{L}^2 -local minimizer.

Example 3. In (40), put

$$(41a) \quad r(t) = \begin{cases} 1 - 2t, & t \in [0, \frac{1}{2}], \\ (1 - 2t)(4t - 1), & t \in (\frac{1}{2}, 1]; \end{cases}$$

$$(41b) \quad S(t) = 4t - 1;$$

$$(41c) \quad Q(t) = \begin{cases} -\frac{1}{e^2}, & t \in [0, \frac{1}{2}], \\ 0, & t \in (\frac{1}{2}, 1]. \end{cases}$$

For simplicity, let

$$(41d) \quad A(t) = 1$$

and

$$(41e) \quad B(t) = 1.$$

(However, any nonnegative bounded measurable $A(t)$ and $B(t)$ with the suitable $Q(t)$ will still work.)

We claim that u^* is an \mathcal{L}^∞ -local minimizer, but not an \mathcal{L}^2 -local minimizer. The latter statement is clear, since $S(u^*)(t) = S(t)$ violates the necessary condition (23) of

\mathcal{L}^2 -local optimality. In fact, from this point of view, there is no \mathcal{L}^2 -local minimizer for (40), (41). To see the former claim, let x^* and ψ^* be the state and costate functions corresponding to u^* , respectively. Note that $x^*(t) = 0$ for t in $[0, \frac{1}{2}]$; therefore

$$(42a) \quad Q(t)x^*(t) = 0, \quad t \in [0, 1],$$

and (7), (8), and (41a) yield

$$(42b) \quad \psi^*(t) = 0.$$

Now let us check the sufficient conditions for \mathcal{L}^∞ -local optimality in Theorem 4. Conditions (6)–(13), (40), (41), (42a), and (42b) give

$$(43a) \quad \nabla J(u^*)(t) = \begin{cases} 1 - 2t, & t \in [0, \frac{1}{2}], \\ 0, & t \in (\frac{1}{2}, 1]; \end{cases}$$

$$(43b) \quad S(u^*)(t) = 4t - 1 \geq 1, \quad t \in (\frac{1}{2}, 1];$$

$$(43c) \quad K(u^*)(t, s) = 0, \quad t \in (\frac{1}{2}, 1] \text{ or } s \in (\frac{1}{2}, 1];$$

$$(43d) \quad \alpha(u^*) = [0, \frac{1}{2}];$$

and

$$(43e) \quad T(u^*) = \{w \in \mathcal{L}^2(0, 1) : w(t) = 0 \text{ a.e. in } \alpha(u^*)\}.$$

Conditions (43) show that the formal Kuhn–Tucker conditions (25) are satisfied at u^* . In addition, $\alpha(u^*)$ is closed, and $S(u^*)(t)$ is continuous at the frontier of $\alpha(u^*)$ in $[0, 1]$. Therefore the second-order sufficient conditions for \mathcal{L}^∞ -local optimality hold at u^* .

The preceding conclusions can also be established directly as follows. Let $v \in \mathcal{L}^\infty(0, 1)$, $u^* + v \in \Omega^+$, and let $x_v(t)$ be the unique solution of (40b) with u being $u^* + v$. Then

$$(44a) \quad 0 \leq x_v(t) \leq e \int_0^t [u^*(\tau) + v(\tau)] d\tau.$$

In addition, assume that $\|v\|_\infty < 1$. Then (42a), (44a), and Cauchy inequality yield

$$\begin{aligned} J(u^* + v) - J(u^*) &= \int_0^1 \left[r(t)v(t) + S(t)u^*(t)v(t) + \frac{1}{2}S(t)v^2(t) + \frac{1}{2}Q(t)x_v^2(t) \right] dt \\ &= \int_0^{1/2} (1 - 2t)v(t)dt + \frac{1}{2} \int_0^1 (4t - 1)v^2(t)dt - \frac{1}{2e^2} \int_0^{1/2} x_v^2(t)dt \\ &\geq \int_0^{1/2} \left[(1 - 2t) + \frac{1}{2}(4t - 1) \right] v^2(t)dt + \frac{1}{2} \int_{1/2}^1 v^2(t)dt \\ &\quad - \frac{1}{4e^2} \int_0^{1/2} e^2 \int_0^{1/2} v^2(t)dt \\ &= \frac{1}{2} \int_0^{1/2} v^2(t)dt + \frac{1}{2} \int_{1/2}^1 v^2(t)dt - \frac{1}{8} \int_0^{1/2} v^2(t)dt \\ &\geq \frac{1}{4} \int_0^1 v^2(t)dt. \end{aligned}$$

Therefore u^* is a strict \mathcal{L}^∞ -local minimizer of J on Ω^+ (compare (44b) with (30)). On the other hand, a similar estimate shows that for $k > 2$, $J(u^* + v_n) - J(u^*)$ is eventually negative on the sequence $v_n = k \cdot (\text{characteristic function of } [0, 1/n])$, and therefore u^* is not an \mathcal{L}^2 -local minimizer.

Finally, we note that J cannot be convex on Ω^+ , since u^* satisfies the first-order necessary conditions for optimality in (25), yet u^* is not a global minimizer of J in Ω^+ .

As noted above, u^* cannot be an \mathcal{L}^2 -local minimizer because $S(t)$ is negative on a subset of $[0, 1]$ with positive measure. In the next example, $S(t)$ and $r(t)$ are adjusted so that u^* is \mathcal{L}^2 -locally optimal, while J remains nonconvex on Ω^+ .

Example 4. Let the constant c be in $(0, c_o)$, where

$$(45a) \quad c_o = 2e^{-2}(e - e^{1/2})(1 - e^{-1/4})^2 > 0.$$

In (38), put

$$(45b) \quad r(t) = \begin{cases} 1 - 2t, & t \in [0, \frac{1}{2}], \\ c(1 - 2t), & t \in (\frac{1}{2}, 1]; \end{cases}$$

$$(45c) \quad S(t) = c, \quad t \in [0, 1],$$

while $u^*(t)$, $Q(t)$, $A(t)$, and $B(t)$ remain the same as defined in (40d) and Example 3. As in Example 3, we can find that the formal Kuhn–Tucker conditions (25) hold at u^* , $\alpha(u^*)$ is closed, and $S(u^*)(t) = c$ is continuous at the frontier of $\alpha(u^*)$ in $[0, 1]$. Moreover,

$$(45d) \quad K(u^*)(t, s) = \int_{\max(t, s)}^1 e^{\tau-t} Q(\tau) e^{\tau-s} d\tau \in \mathcal{L}^\infty([0, 1] \times [0, 1]),$$

and $S(u^*)(t)$ is positive and bounded away from 0 in $[0, 1]$. By Theorem 4, u^* is an \mathcal{L}^2 -local minimizer of J in Ω^+ . This fact can also be easily shown with a special version of the final estimates in the proof of Theorem 4.

We now show that J is nonconvex on Ω^+ . Put

$$(46a) \quad v_o(t) = \begin{cases} 1, & t \in [0, \frac{1}{4}]; \\ 0, & t \in (\frac{1}{4}, 1]. \end{cases}$$

Then $u^* + v_o \in \Omega^+$, and, for $(t, s) \in [0, \frac{1}{4}] \times [0, \frac{1}{4}]$, (41c) and (45d) yield

$$(46b) \quad \begin{aligned} K(u^*)(t, s) &= -e^{-(2+t+s)} \int_{\max(t, s)}^{1/2} e^{2\tau} d\tau \\ &\leq -e^{-(2+t+s)} \int_{1/4}^{1/2} e^{2\tau} d\tau \\ &= -\frac{1}{2} e^{-(2+t+s)} (e - e^{1/2}). \end{aligned}$$

So (2c), (2d), (45), and (46) yield

$$\begin{aligned} d^2 J(u^*; v_o, v_o) &\leq \int_0^{1/4} c \, dt + \int_0^{1/4} \int_0^{1/4} \left(-\frac{1}{2}\right) e^{-(2+t+s)} (e - e^{1/2}) \, ds \, dt \\ &= \frac{1}{4}c - \frac{1}{2}e^{-2}(e - e^{1/2})(1 - e^{1/4})^2 \\ &= \frac{1}{4}(c - c_o) < 0. \end{aligned}$$

Therefore the restriction of J to Ω^+ is nonconvex. Furthermore, since v_o is a direction of recession at u^* in Ω^+ , we can see that u^* is not a global minimizer; in fact, $\inf_{\Omega^+} J = -\infty$.

In the third example below, an optimization problem is constructed in such a way that the \mathcal{L}^2 -local second-order sufficient conditions in Theorem 4 hold at the global minimizer u^* of a nonconvex quadratic objective function J on Ω^+ . Compare this with the previous examples and recall that, for unconstrained minimization problems with quadratic objective functions J , the standard second-order sufficient conditions imply that J is convex, and a global minimizer exists only if J is convex.

Example 5. Consider

$$\begin{aligned} \min J(u) &= \int_0^1 \left\{ r(t)u(t) + \frac{1}{2}[u(t) - u^*(t)]^2 \right\} dt \\ (47a) \quad &+ \frac{1}{2} \int_0^1 \int_0^1 K(t, s)[u(s) - u^*(s)][u(t) - u^*(t)] \, ds \, dt, \\ u \in \Omega^+ &= \{w \in \mathcal{L}^2(0, 1), \quad w(t) \geq 0 \quad \text{a.e. in } [0, 1]\}, \end{aligned}$$

where

$$(47b) \quad r(t) = \begin{cases} 3, & t \in [0, \frac{1}{2}], \\ 0, & t \in (\frac{1}{2}, 1]; \end{cases}$$

$$(47c) \quad K(t, s) = \begin{cases} 0, & (t, s) \in ([0, \frac{1}{2}] \times [0, \frac{1}{2}]) \cup ([\frac{1}{2}, 1] \times [\frac{1}{2}, 1]), \\ 5, & \text{otherwise;} \end{cases}$$

and

$$(47d) \quad u^*(t) = \begin{cases} 0, & t \in [0, \frac{1}{2}], \\ 1, & t \in (\frac{1}{2}, 1]. \end{cases}$$

It is easy to see that for $u \in \Omega^+$ and $v \in \mathcal{L}^2(0, 1)$,

$$(48a) \quad d^1 J(u; v) = \int_0^1 \left[r(t) + u(t) - u^*(t) + \int_0^1 K(t, s)(u(s) - u^*(s)) \, ds \right] v(t) \, dt,$$

$$(48b) \quad d^2 J(u; v, v) = \int_0^1 v^2(t) \, dt + \int_0^1 \int_0^1 K(t, s)v(s)v(t) \, ds \, dt,$$

and

$$(48c) \quad \alpha(u^*) = [0, \tfrac{1}{2}],$$

$$(48d) \quad T(u^*) = \{w \in \mathcal{L}^2(0, 1), \quad w(t) = 0 \quad \text{a.e. in } [0, \tfrac{1}{2}]\}.$$

Compare (48a), (48b) with (2) to find that

$$(49a) \quad \nabla J(u^*)(t) = r(t),$$

$$(49b) \quad S(u)(t) = 1, \quad \forall u \in \mathcal{L}^2(0, 1),$$

and

$$(49c) \quad K(u)(t, s) = K(t, s), \quad \forall u \in \mathcal{L}^2(0, 1).$$

By (48) and (49), conditions (3) are satisfied and the formal Kuhn–Tucker conditions (25) hold at u^* ; moreover, $\alpha(u^*)$ is closed and $S(u^*)(t)$ is continuous, positive and bounded away from zero in $[0, 1]$. Therefore the \mathcal{L}^2 -local sufficient conditions in Theorem 4 hold at u^* .

In fact, u^* is the global minimizer of J in Ω^+ . To see this, let u be any element in Ω^+ . Then

$$(50a) \quad u(t) - u^*(t) \geq 0, \quad t \in [0, \tfrac{1}{2}],$$

$$(50b) \quad u(t) - u^*(t) \geq -1, \quad t \in (\tfrac{1}{2}, 1],$$

and (47)–(50) give

$$\begin{aligned} J(u) - J(u^*) &= d^1 J(u^*; u - u^*) + \frac{1}{2} d^2 J(u^*; u - u^*, u - u^*) \\ &= \int_0^{1/2} 3[(u(t) - u^*(t))] dt + \frac{1}{2} \int_0^1 [(u(t) - u^*(t))^2] dt \\ &\quad + 5 \int_0^{1/2} [(u(t) - u^*(t))] dt \int_{1/2}^1 [(u(s) - u^*(s))] ds \\ &\geq (3 - \tfrac{5}{2}) \int_0^{1/2} [(u(t) - u^*(t))] dt + \frac{1}{2} \int_0^1 [(u(t) - u^*(t))^2] dt \\ &\geq \frac{1}{2} \int_0^1 [(u(t) - u^*(t))^2] dt \geq 0. \end{aligned}$$

Therefore u^* is globally optimal.

Finally, we show that the restriction of J to Ω^+ is not convex. Put

$$v_o(t) = \begin{cases} 1, & t \in [0, \tfrac{1}{2}], \\ -1, & t \in (\tfrac{1}{2}, 1]. \end{cases}$$

Then $u^* + v_o \in \Omega^+$, and

$$\begin{aligned} d^2 J(u^*; v_o, v_o) &= \int_0^1 v_o^2(t) dt + \int_0^1 \int_0^1 K(t, s) v_o(s) v_o(t) ds dt \\ &= 1 + 2 \cdot 5 \int_0^{1/2} 1 dt \int_{1/2}^1 (-1) ds \\ &= 1 - 10 \cdot \left(\frac{1}{2}\right)^2 = 1 - \frac{5}{2} < 0. \end{aligned}$$

So J is nonconvex on Ω^+ .

Note Added in Proof. Two additional papers of related interest have come to our attention and should be mentioned here. Reference [13] formulates general Banach space sufficient conditions that insure local quadratic growth estimates in one norm within small neighborhoods described by a second norm (as in Lemma 1). In the present context, the first- and second-order coercivity conditions in (29a) and (29b) imply a coercivity condition similar to (3.5) in [13] on the intersection of a "conical neighborhood" of T_α with Ω^+ . Reference [13] does not establish a counterpart of the coercivity extension process in Lemma 2, or an \mathcal{L}^2 -local optimality result like the one in Theorem 4; moreover, the \mathcal{L}^∞ -local optimality result for control problems in Theorem 5.2 of [13] requires the strict Legendre-Clebsch condition on the entire interval $[0, 1]$ (cf. Theorems 2 and 4). On the other hand, the general control problem sufficiency analysis in [13] has a wide scope, and relates the required coercivity property to Legendre-Clebsch and Jacobian disconjugacy conditions (cf. Note 8). Finally, Lemma 8 in [14] is a single-norm variant of Lemma 1 for general abstract nonlinear programs, with coercivity hypotheses that are stronger than (29) in the present setting.

REFERENCES

- [1] E. POLAK, *Computational Methods in Optimization*, Academic Press, New York, 1971.
- [2] E. POLAK AND L. HE, *Rate preserving discretization strategies for semi-infinite programming and optimal control*, Electronics Research Laboratory Memorandum, University of California, Berkeley, CA, 1990.
- [3] 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.
- [4] J. C. DUNN, *Extremal types for certain \mathcal{L}^p -minimization problems and associated large scale nonlinear programs*, Appl. Math. Optim., 10 (1983), pp. 303–335.
- [5] ———, *Diagonally modified conditional gradient methods for input constrained optimal control problems*, SIAM J. Control Optim., 24 (1986), pp. 1177–1191.
- [6] J. C. DUNN AND E. SACHS, *The effects of perturbations on the convergence rates of optimization algorithms*, Appl. Math. Optim., 10 (1983), pp. 143–157.
- [7] W. HAGER, *Multiplier methods for nonlinear optimal control*, SIAM J. Numer. Anal., 27 (1990), pp. 1061–1080.
- [8] G. J. ZALMAI, *Generalized sufficiency criteria in continuous-time programming with applications to a class of variational-type inequalities*, J. Math. Anal. Appl., 153 (1990), pp. 331–355.
- [9] R. P. BOAS, *A Primer of Real Functions*, 3rd ed., Carus Mathematical Monographs, 13, The Mathematical Association of America, Washington, DC, 1981.
- [10] H. L. ROYDEN, *Real Analysis*, 2nd ed., Macmillan, New York, 1968.
- [11] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, Krieger Publishing, Huntington, NY, 1980.
- [12] V. M. ALEKSEEV, V. M. TIKHOMIROV, AND S. V. FOMIN, *Optimal Control*, Plenum Publishing, New York, 1987.
- [13] H. MAURER, *First- and second-order sufficient optimality conditions in mathematical programming and optimal control*, Mathematical Programming Study, 14 (1981), pp. 163–177.

- [14] A. L. DONTCHEV AND W. W. HAGER, *Lipschitzian stability in nonlinear control and optimization*, SIAM J. Control Optim., to appear.

EXTREME POINTS FOR LINEAR OPTIMAL CONTROL PROBLEMS WITH DIAGONAL STRUCTURE*

EDWARD J. ANDERSON[†] AND ANDREW B. PHILPOTT[‡]

Abstract. This paper discusses the extreme points of the feasible set for a certain type of optimal control problem with constraints on both the state and control variables. The problem is posed as a continuous-time linear program in the space of bounded measurable functions. The extreme points can be characterized by a certain full rank condition.

Key words. continuous linear program, extreme points, linear optimal control, state constraints

AMS(MOS) subject classifications. 90C45, 49N05

1. Introduction. In this paper, we consider linear optimal control problems with linear constraints on the control variables and box constraints on the state variables. One linear optimal control problem (LOC) of this type can be formulated as follows:

$$\begin{aligned} \text{LOC:} \quad & \text{minimize} \quad \int_0^T (c_1(t)x(t) + c_2(t)u(t))dt, \\ & \text{subject to} \quad (d/dt)x(t) = A(t)x(t) + B(t)u(t) + g(t), \\ & \quad \quad \quad H(t)u(t) \leq b(t), \\ & \quad \quad \quad x(0) = x_0, \quad 0 \leq x(t) \leq d(t), \\ & \quad \quad \quad u(t) \geq 0, \quad t \in [0, T], \end{aligned}$$

where $c_1(t), c_2(t), g(t), b(t)$, and $d(t)$ are given vectors; $A(t), B(t)$ and $H(t)$ are given matrices; and x_0 is a given initial state. Our aim is to identify a set of solutions amongst which the optimal solution can be guaranteed to lie.

The approach we take to this problem is motivated by consideration of the more general continuous-time linear program, which is usually formulated as follows:

$$\begin{aligned} \text{CLP:} \quad & \text{minimize} \quad \int_0^T c(t)x(t)dt, \\ & \text{subject to} \quad B(t)x(t) + \int_0^t K(t,s)x(s)ds = b(t), \\ & \quad \quad \quad x(t) \geq 0, \quad t \in [0, T], \end{aligned}$$

where $c(t)$ and $b(t)$ are given vectors (functions of time), and $B(t)$ and $K(t, s)$ are given matrices. Continuous-time linear programs were first considered in the literature by Bellman [4], who coined the term *bottleneck problem* to describe them.

In this framework, it is natural to seek a simplex-like algorithm for the solution of the problem. A number of authors, returning to the pioneering work of Lehman [8], have attempted to develop the theory of linear programming for the general problem CLP. There are three ingredients in classical linear programming that we might hope to duplicate in the continuous-time context: a theory of the relationship between CLP and its dual problem; the identification of basic solutions (which are just extreme points of the feasible set); and a pivot step to move from any suboptimal basic solution to a better one. Of these three elements, the greatest amount of work has been done

* Received by the editors May 16, 1990; accepted for publication (in revised form) August 23, 1991.

[†] University of Cambridge, Judge Institute of Management Studies, Mill Lane, Cambridge CB2 1RX, England.

[‡] Department of Engineering Science, University of Auckland, New Zealand.

on duality theory (e.g., Tyndall [12], Grinold [6]), but many authors have attempted progress toward an algorithm for CLP (e.g., Drews [5], Hartberger [7], and Segers [11]). The most ambitious attempt to develop an algorithm for the problem is represented by the work of Perold [9] and Anstreicher [3]. In some aspects, the results have been disappointing; the theory that emerges is highly complex, and there remain substantial difficulties that make an implementation of the method in any automatic way unlikely to be successful.

A characterization of the extreme points of the feasible set is a prerequisite for the development of a simplex-like algorithm for CLP. In this paper, we concentrate on exploring the structure of the extreme points for problems of the form LOC. The more general problem of describing the extreme points of CLP has been addressed by Perold [10]. He considers the problem CLP in which the matrices B and K are constant, and so the constraints take the form

$$Bx(t) + \int_0^t Kx(s)ds = b(t), \quad x(t) \geq 0, \quad t \in [0, T].$$

Moreover, he considers solutions $x(t)$ that are right analytic functions. The effect of this is to divide the interval $[0, T]$ into a (possibly infinite) series of intervals, in each of which $x(t)$ is analytic (for details, refer to [10]). In each interval I_j , say, we can define the *basis* β_j as the set of indices of components of $x(t)$ that are nonzero on this interval. Within this framework, Perold has shown that a necessary and sufficient condition for a right analytic solution $x(t)$ to be an extreme point of the feasible set is that, for each interval I_j , there is some scalar μ such that the columns of $\mu B + K$ indexed by β_j are linearly independent.

There are two weaknesses in Perold's result. First, there is no guarantee that there will be an extreme-point optimal solution that is right analytic, even if the functions appearing in the problem formulation are all analytic. From the point of view of computation, this is not a restriction. If we wish to implement a continuous-time simplex algorithm for CLP, we must, in any case, work with feasible solutions having only a finite number of points at which the basis set changes, which will therefore be analytic within these constant basis intervals. If the optimal solution does not have a finite number of constant basis intervals, then it cannot be found by such an algorithm, and the best that we could hope for is that the algorithm produces a sequence that converges to the optimal solution. Nevertheless, from a theoretical point of view, it is desirable to know that the optimal solution exists and is an extreme point, which is guaranteed if we work in the space of essentially bounded measurable functions (so that all the functions appearing in the problem statement are bounded and measurable), and, in addition, there is some bound on the feasible region. This is the framework we adopt in this paper.

A second weakness in Perold's result is the restriction to problems in which the constraint matrices are constant. Perold gives examples in [10] to show that, in some circumstances, time dependence in the constraint matrices can lead to a failure of the basis characterization of extreme points. Nevertheless, as we show below, there are instances of CLP having time-dependent constraint matrices for which a characterization of extreme points is straightforward.

The class of continuous linear programs that we consider is a special case of the problem LOC in which the matrix A is diagonal. Note that the problem formulation LOC already separates the problem variables into two sets: the state variables and the control variables, whereas there is no such distinction for CLP. We make the

additional assumption that the state feedback operates in a univariate fashion, so that the rate of change in the state variable x_i is affected only by the controls u and by the value of x_i . The approach we take here is similar to that used by Anderson, Nash, and Perold [1] for the even more restricted class of problems called separated continuous linear programs. These are obtained from LOC by setting the matrix A to zero, removing the time dependence from the matrix B , and removing the upper bound constraints on the state variables.

Anderson and Philpott [2] have shown how the characterization of extreme points for separated continuous linear programs can be used in the development of an algorithm to solve the network program formulated in continuous time with the possibility of storage at the nodes of the network. The problem formulation that we consider here can also be applied in a network environment when there are “gains” operating at the nodes of the network. This might occur with negative gains if the commodity in question was subject to systematic losses over time. Alternatively, as an example of a problem with positive gains, we could consider the transfer of money between different interest-yielding locations, so that money stored at the nodes of the network is steadily increased.

2. A characterization of extreme points for LOC. In this section, we give a characterization of the extreme points of LOC in the case where A is diagonal. We begin by restating the constraints that determine the feasible region of LOC. It is convenient to work with the integral form of the dynamics and to introduce a slack variable into the constraint on the control variables. We obtain

$$(1) \quad x(t) = \int_0^t (A(s)x(s) + B(s)u(s)) ds + a(t),$$

$$(2) \quad H(t)u(t) + w(t) = b(t),$$

$$(3) \quad \begin{aligned} 0 &\leq x(t) \leq d(t), \\ u(t), w(t) &\geq 0, \quad t \in [0, T]. \end{aligned}$$

Here the function $a(t)$ is continuous and takes the place, in terms of the previous notation, of $x_0 + \int_0^t g(s) ds$. The components of b are bounded measurable functions defined on $[0, T]$, and those of d are continuous functions on $[0, T]$. The matrices $A(t)$, $B(t)$, and $H(t)$ have components that are bounded measurable functions defined on $[0, T]$, with $A(t)$ being $n \times n$, where n is the number of state variables, and $B(t)$ and $H(t)$ have dimensions determined by those of u and b .

It is clear from (1) that any choice of controls $u(s)$, $s \in [0, t]$ will uniquely determine a vector $x(t)$ of state variables, continuous in t . If $u(t)$ and $w(t)$ are nonnegative, satisfy (2) for all $t \in [0, T]$, and generate state variables $x(t)$ satisfying (3) for all $t \in [0, T]$, then we say that (u, x, w) is *feasible*. The set of all feasible (u, x, w) is denoted by \mathcal{F} .

Our first concern is to establish the existence of extreme-point optimal solutions for LOC. To do this, we must work within the space L_∞ of essentially bounded measurable functions in which functions that differ on a set of measure zero are identified. We write F for the set of points in L_∞ that correspond to a point in \mathcal{F} . The set F can be viewed as the set of essentially nonnegative (u, x, w) in L_∞ that satisfy (1)–(3) almost everywhere.

THEOREM 1. *If the constraints (2) bound u for almost all $t \in [0, T]$, then the problem LOC has an optimal solution that is an extreme point of F .*

Proof. This result follows immediately from Theorem 3 of Perold [10] once we observe that, since u is essentially bounded, there is some $M > 0$ with $\|u(t)\| \leq M$, $\|w(t)\| \leq M$, and $\|x(t)\| \leq M$ for almost all $t \in [0, T]$.

Unless stated otherwise, we henceforth assume that the matrix A in (1) is diagonal. If, for such a matrix, we define the *transition matrix* to be

$$\Phi(s, t) = \exp\left[\int_s^t A(\tau) d\tau\right],$$

then the following result, which is well known, gives a formula for determining the change in state variables caused by a change in control variables.

LEMMA 1. *Suppose that (u, x) satisfies (1). If $u' = u + v$ and*

$$x'(t) = x(t) + \int_0^t \Phi(s, t) B(s) v(s) ds,$$

then (u', x') satisfies (1).

We also use the following result in Lebesgue measure theory.

LEMMA 2. *Suppose that f is a nonnegative bounded measurable function on $[0, T]$, and that, for some set P of nonzero measure in $[0, T]$, $f(t) > 0$, $t \in P$. Then*

- (i) *There is some $\epsilon > 0$ and some set P' such that $P' \cap P$ has nonzero measure, and $f(t) > \epsilon$, $t \in P'$;*
- (ii) *If f is continuous, then P' may be taken to be an open subinterval of $[0, T]$.*

Proof. The first part of the lemma follows by letting J_i be the subset of P on which $f(t) > 1/i$. Then the J_i form a nested sequence of measurable sets with $P = \bigcup_{i=1}^{\infty} J_i$. Since P has nonzero measure, so does J_m for some m . Setting $\epsilon = 1/m$ and $P' = J_m$ yields the desired result.

When f is continuous, for each $t \in P$ we may choose $\epsilon(t) > 0$ and an open interval I_t such that $f(t) > \epsilon(t)$, $t \in I_t$. The set $\bigcup_{t \in P} I_t$ contains P , so, for at least one t in this collection, $I_t \cap P$ has nonzero measure. Letting $I = I_t$ and $\epsilon = \epsilon(t)$ establishes the second part of the lemma.

The theorem below gives a characterization of the extreme points of LOC in the case where A is a diagonal matrix. This theorem is an extension of the result in [1]. To state and prove it, we must state some definitions. First, let

$$K(t) = \begin{bmatrix} B(t) & I & 0 \\ H(t) & 0 & I \end{bmatrix}$$

and define the *support* S_f of some bounded measurable vector function f to be the set-valued function of time defined for each $t \in [0, T]$ by

$$S_f(t) = \{k : |f_k(t)| > 0\}.$$

Given a feasible solution (u, x, w) , we denote the set triple $(S_u(t), S_x(t) \cap S_{d-x}(t), S_w(t))$ by $S(t)$.

THEOREM 2. *Suppose that $(u, x, w) \in \mathcal{F}$ and that A is diagonal. Then (u, x, w) corresponds to an extreme point of F if and only if the columns of K indexed by $S(t)$ are linearly independent for almost all $t \in [0, T]$.*

Proof. Our proof is similar to that of Theorem 2 in [1]. The essential idea is to show how a perturbation, which can be added to or subtracted from $u(t)$, is linked

with a vector function of time $\gamma(t)$, which has the property that its support is within $S(t)$ and $K(t)\gamma(t)$ is zero. In the absence of state variables, this would be easy; we could choose as a perturbation any sufficiently small scalar multiple λ of those components of $\gamma(t)$ corresponding to the control variables u . Difficulties arise when a perturbation of u affects the state variables x , since a change to the components of $x(t)$ at some time t_0 affects the values of these components at all future times and may therefore lead to an infeasibility occurring for some $t > t_0$. This is dealt with in two ways. First, we use the continuity of the state variables to define an interval (α, β) in which those state variables that change are not near their bounds, thus avoiding the possibility of infeasibility. Second, we choose different values of the scalar multiple λ in successive subintervals of (α, β) , so that the total effect of the perturbation is to bring the state variables back to their original values at β .

Suppose then that, for some set P of nonzero measure in $[0, T]$, the columns of K indexed by $S(t)$ are linearly dependent for $t \in P$. Since $S(t)$, $t \in P$ can take only a finite number of values, we can assume without loss of generality that $S(t)$ is constant ($= (\bar{S}_u, \bar{S}_x, \bar{S}_w)$, say) on P .

We seek an essentially nonzero function p that can be added to and subtracted from u to give (u', x', w') and (u'', x'', w'') , both of which are in \mathcal{F} . Since the components of x with indices in \bar{S}_x are continuous functions, we can show by Lemma 2 that there is some $\epsilon > 0$, and some interval (α, β) such that $P' = P \cap (\alpha, \beta)$ has nonzero measure and

$$(4) \quad 0 < x_j(t) \pm \epsilon < d_j(t), \quad j \in \bar{S}_x, \quad t \in (\alpha, \beta).$$

If $\bar{S}_x = \emptyset$, then we can take $(\alpha, \beta) = (0, T)$ and some of the steps in the proof below become trivial.

Since the columns of $K(t)$ indexed by $S(t)$ are linearly dependent on P' , there are vectors $p(t)$, $q(t)$, and $r(t)$, (p at least differing from zero for all $t \in P'$), such that

$$(5) \quad \begin{aligned} B(t)p(t) + q(t) &= 0, & t \in P', \\ H(t)p(t) + r(t) &= 0, & t \in P', \end{aligned}$$

$$(6) \quad \begin{aligned} p_k(t) &= 0, & t \notin P' \text{ or } k \notin \bar{S}_u, \\ q_j(t) &= 0, & t \notin P' \text{ or } j \notin \bar{S}_x, \\ r_k(t) &= 0, & t \notin P' \text{ or } k \notin \bar{S}_w. \end{aligned}$$

Again using Lemma 2, P' can be chosen so that, for some $\epsilon' > 0$,

$$u_k(t) \geq \epsilon', \quad t \in P', \quad k \in \bar{S}_u$$

and

$$w_k(t) \geq \epsilon', \quad t \in P', \quad k \in \bar{S}_w.$$

Now choose $t_0 = \alpha < t_1 < t_2 < \cdots < t_{n+1} = \beta$, so that, for each $i = 1, 2, \cdots, n+1$, $P_i = (t_{i-1}, t_i) \cap P'$ has nonzero Lebesgue measure. For $i = 1, 2, \cdots, n+1$, define $\theta^{(i)}$ by

$$\theta^{(i)} = \int_{P_i} \Phi(s, t_{n+1}) q(s) ds.$$

Since each $\theta^{(i)}$ has only n components, there exist $\lambda_1, \lambda_2, \dots, \lambda_{n+1} \in [-1, 1]$ not all zero with

$$\sum_{i=1}^{n+1} \lambda_i \theta^{(i)} = 0.$$

If we define

$$\begin{aligned} u'(t) &= u(t) + \delta \sum_{i=1}^{n+1} \lambda_i p(t) \chi_{P_i}, & u''(t) &= u(t) - \delta \sum_{i=1}^{n+1} \lambda_i p(t) \chi_{P_i}, \\ w'(t) &= w(t) + \delta \sum_{i=1}^{n+1} \lambda_i r(t) \chi_{P_i}, & w''(t) &= w(t) - \delta \sum_{i=1}^{n+1} \lambda_i r(t) \chi_{P_i}, \end{aligned}$$

where χ_{P_i} denotes the characteristic function of P_i , then $H(t)u'(t) + w'(t) = b(t)$, and $H(t)u''(t) + w''(t) = b(t)$, and, by choosing $\delta > 0$ small enough, we can guarantee that, for every $t \in [0, T]$,

$$u'(t) \geq 0, \quad u''(t) \geq 0,$$

$$w'(t) \geq 0, \quad w''(t) \geq 0.$$

It remains to show that $\delta > 0$ can be chosen so that $u'(t)$ and $u''(t)$ will generate respective state variables $x'(t)$ and $x''(t)$ satisfying (3). We proceed to show that $x'(t)$ and $x''(t)$ differ from $x(t)$ on (α, β) by less than ϵ given by (4), and are the same as $x(t)$ outside this interval. We give the argument for $x'(t)$ only, for which we obtain, by Lemma 1,

$$x'(t) = x(t) + \delta \sum_{i=1}^{n+1} \lambda_i \int_{(0,t) \cap P_i} \Phi(s, t) B(s) p(s) ds.$$

Since $\int_{(0,t) \cap P_i} \Phi(s, t) B(s) p(s) ds$ is a bounded function of t on $[0, T]$, $\delta > 0$ can be chosen so that $x'_j(t)$ differs from $x_j(t)$ by less than ϵ , for $t \in (\alpha, \beta)$ and $j \in \bar{S}_x$. Because $A(t)$ is diagonal by assumption, so is $\Phi(s, t)$; thus the j th component of $\int_{(0,t) \cap P_i} \Phi(s, t) B(s) p(s) ds$ vanishes when $(B(t)p(t))_j = 0$, for almost all $t \in P'$. In particular, this will occur for $j \notin \bar{S}_x$ by virtue of (5) and (6), and so every component of $x'(t)$ satisfies (3), for $t \in (t_0, t_{n+1})$.

Finally, it is clear that $x'(t) = x(t)$, $t \leq t_0$, and at t_{n+1} we have

$$x'(t_{n+1}) = x(t_{n+1}) + \delta \sum_{i=1}^{n+1} \lambda_i \int_{(0, t_{n+1}) \cap P_i} \Phi(s, t_{n+1}) B(s) p(s) ds,$$

which by (5) and the definition of the multipliers λ_i gives

$$x'(t_{n+1}) = x(t_{n+1}) - \delta \sum_{i=1}^{n+1} \lambda_i \theta^{(i)} = x(t_{n+1}).$$

Thus, since $u'(t) = u(t)$, $t \geq t_{n+1}$, it follows that $x'(t) = x(t)$, $t \geq t_{n+1}$, which shows that $(u', x', w') \in \mathcal{F}$. Repeating this argument for $u''(t)$ completes the first half of the

proof, since this gives (u, x, w) as a convex combination of (u', x', w') and (u'', x'', w'') , which correspond to distinct members of F .

For the converse, suppose that (u, x, w) is not an extreme point of F , so that there exists (u', x', w') and (u'', x'', w'') both in \mathcal{F} , with $(u', x', w') \neq (u'', x'', w'')$ on a set of nonzero measure in $[0, T]$, and

$$(u, x, w) = \frac{1}{2}(u', x', w') + \frac{1}{2}(u'', x'', w'').$$

Let

$$v(t) = u'(t) - u(t), \quad y(t) = -B(t)v(t), \quad \text{and} \quad z(t) = w'(t) - w(t).$$

It is easily verified from (2) that $H(t)v(t) + z(t) = 0$, and so

$$K(t) \begin{pmatrix} v(t) \\ y(t) \\ z(t) \end{pmatrix} = 0.$$

It is clear from the definition of v and z that $v_k \neq 0$ only if $u_k > 0$, and $z_k \neq 0$ only if $w_k > 0$. So it only remains to show that $y_j(t) \neq 0$ implies that $0 < x_j(t) < d_j(t)$.

Consider the function $e(t) = x'(t) - x(t)$. It follows by differentiating (1) that, for almost all $t \in [0, T]$,

$$(d/dt)e(t) = A(t)e(t) - y(t).$$

Consider the j th component of x . Let M_j be the set on which either $x_j(t) = 0$ or $x_j(t) = d_j(t)$. We proceed to show that $y_j(t) = 0$, for almost every t in M_j . Since both x' and $x''(t)$ are feasible, $e_j(t) = 0$ on M_j . Let M'_j be the subset of M_j where the derivative of e_j is nonzero. Now, for each $\tau \in M'_j$, there is some $\delta > 0$ so that, for $0 < |h| < \delta$,

$$\frac{e_j(\tau + h) - e_j(\tau)}{h}$$

is bounded away from zero. Thus for each τ there exists δ_τ with no other element of M_j within a distance δ_τ of τ . Since $M'_j \subseteq M_j$, the interval $(\tau - \delta_\tau, \tau)$ contains no element of M'_j , and so we can establish a 1-1 correspondence between M'_j and a subset of the rational numbers (for each τ choose a rational number in the interval $(\tau - \delta_\tau, \tau)$). It follows that M'_j is countable and thus has measure zero. Therefore, since $A(t)$ is diagonal, for almost all $t \in M_j$, $A_{j,j}(t)e_j(t) - y_j(t) = 0$, and hence $y_j(t) = 0$. Thus we have shown that $v(t)$, $y(t)$, and $z(t)$ determine a linear dependence amongst the columns of $K(t)$ indexed by $S(t)$, except when $t \in \bigcup M'_j$, a set of measure zero.

3. Discussion. Although at first sight the characterization we have given might appear to be different to the condition given by Perold [10], we can easily show that they are equivalent. When translated into the terms we have used here, Perold's result can be expressed by saying that a (right analytic) solution is an extreme point if and only if, for each interval J on which the solution is analytic, there is some μ for which the columns of the matrix

$$\mu \begin{bmatrix} 0 & I & 0 \\ H & 0 & I \end{bmatrix} + \begin{bmatrix} B & A & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

indexed by nonzero elements of u , x , and w on J are linearly independent. We have dropped the time dependence from A , B , and H since Perold's result is for the time invariant case. Thus, by Perold's condition, a feasible right analytic solution is *not* an extreme point of F if and only if, for some interval J and every μ , there exist vectors p , q , and r with

$$Bp + (\mu I + A)q = 0, \quad \mu Hp + \mu r = 0$$

and with the supports of p , q , and r lying within the supports on J of u , x , and w , respectively. This can be seen to be equivalent to the condition we have given above (in the case that A is diagonal) by introducing a variable \hat{q} with

$$\hat{q}_i = (\mu + A_{i,i})q_i.$$

Then the vector with components p , \hat{q} , r can be used to demonstrate that (u, x, w) is not extreme by Theorem 2 above.

It is interesting to speculate on whether a characterization of extreme points in an L_∞ setting might be possible for some more general formulation than the one we have given. A major source of difficulty in this respect is the fact that bounded measurable functions can have very poorly behaved support. As an example of pathological support, consider the following construction. Let q_1, q_2, \dots be an enumeration of the rationals in $(0, 1)$ and define

$$Q = (0, 1) \cap \bigcup_{i=1}^{i=\infty} (q_i - 2^{-i-2}, q_i + 2^{-i-2}).$$

Now let $P = [0, 1] \setminus Q$. It is clear that the measure of P is at least $\frac{1}{2}$, and yet P contains no open interval. In general, the possibility of encountering such a set means that we cannot remove the diagonal restriction on A , as the following example shows.

Example 1. In LOC, let $H = 0$ and let

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Define P as above and let

$$a_1(t) = -t^2/2 - \int_{(0,t) \cap P} ds, \quad a_2(t) = 0.$$

Let the only inequality constraints be nonnegativity constraints on u and x . The constraints of the problem, in derivative form, are thus as follows:

$$\begin{aligned} (d/dt)x_1(t) &= x_1(t) + x_2(t) + u_1(t) - t - \chi_P(t), \\ (d/dt)x_2(t) &= u_2(t), \\ x_1, x_2, u_1, u_2 &\geq 0, \quad t \in [0, 1]. \end{aligned}$$

A feasible solution for this problem is

$$u_1 = \chi_P, \quad u_2 = 1, \quad x_1 = 0, \quad x_2 = t,$$

and for this solution $S_x(t) = \{2\}$ and $S_u(t) = \{1, 2\}$, for $t \in P$, giving the columns of the matrix $[B \quad \mu I + A]$ indexed by $S(t)$, for $t \in P$ as

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & \mu \end{bmatrix}.$$

Since these are linearly dependent for every μ , the straightforward application of Perold's condition would lead to the conclusion that this solution was not extreme. In fact, it is an extreme point, since, if

$$(u, x) = \frac{1}{2}(u', x') + \frac{1}{2}(u'', x''),$$

then $x'_1 = x''_1 = x_1 = 0$, and $u'_1 = u''_1 = u_1 = 0$, $t \in Q = (0, 1) \setminus P$. Thus

$$x'_2 = -u'_1 - a_1 = -a_1(t), \quad t \in Q,$$

whence $x'_2 = t$, $t \in Q$. Similarly $x''_2 = t$, $t \in Q$. Now x'_2 and x''_2 are continuous functions, and Q contains every rational number in $(0, 1)$, so

$$x''_2(t) = x'_2(t) = x_2(t) = t, \quad t \in [0, 1].$$

It follows immediately that

$$u''_2(t) = u'_2(t) = u_2(t) = 1, \quad \text{a.e. } t \in [0, 1],$$

which shows that (u, x) is extreme.

We note that it appears that examples like this can only occur when the support of some variable has extremely poor behaviour. For example, if P is chosen so as to contain any open interval, then it is easy to construct along the lines of the proof of Theorem 2 nontrivial feasible perturbations (u', x') and (u'', x'') of (u, x) (which return the state variables to their original values at the end of the interval), thereby demonstrating that (u, x) is not extreme. Observe finally that by a suitable change of variables, the above example can also be used to show that a characterization of extreme points for LOC along the lines of Perold's condition is, in general, not possible if the box constraints on the state variables are replaced by general linear inequalities, even if the matrix A is assumed to be diagonal.

REFERENCES

- [1] 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.
- [2] E. J. ANDERSON AND A. B. PHILPOTT, *A continuous-time network simplex algorithm*, Networks, 19 (1989), pp. 395–425.
- [3] K. M. ANSTREICHER, *Generation of feasible directions in continuous time linear programming*, Tech. Report SOL 83-18, Department of Operations Research, Stanford University, Stanford, CA, 1983.
- [4] R. E. BELLMAN, *Bottleneck problems and dynamic programming*, Proc. Nat. Acad. Sci., 39 (1953).
- [5] W. P. DREWS, *A simplex-like algorithm for continuous-time linear optimal control*, in Optimization Methods in Resource Allocation, R. W. Cottle and J. Krarup, eds., Crane Russak, New York, 1974.
- [6] R. GRINOLD, *Symmetric duality for a class of continuous linear programming problems*, SIAM J. Appl. Math., 18 (1970), pp. 84–97.
- [7] R. J. HARTBERGER, *Representation extended to continuous time*, in Optimization Methods in Resource Allocation, R. W. Cottle and J. Krarup, eds., Crane Russak, New York, 1974.

- [8] R. S. LEHMAN, *On the continuous simplex method*, RAND Research Memorandum, RM 1386, Santa Monica, CA, 1954.
- [9] A. F. PEROLD, *Fundamentals of a continuous-time simplex method*, Report SOL 78-26, Department of Operations Research, Stanford University, Stanford, CA, 1978.
- [10] ———, *Extreme points and basic feasible solutions in continuous time linear programming*, SIAM J. Control Optim., 19 (1981), pp. 52–63.
- [11] R. G. SEGERS, *A generalized function setting for dynamic optimal control problems*, in Optimization Methods in Resource Allocation, R. W. Cottle and J. Krarup, eds., Crane Russak, New York, 1974.
- [12] W. F. TYNDALL, *An extended duality theory for continuous linear programming problems*, SIAM J. Appl. Math., 15 (1965), pp. 644–666.

A FINITE FUEL STOCHASTIC CONTROL PROBLEM ON A FINITE TIME HORIZON*

A. P. N. WEERASINGHE†

Abstract. A player starts at x in $[0, a)$ with an initial amount of fuel $y > 0$ and seeks to reach the goal a by time t_0 before spending all the fuel. Fuel is spent by the player at zero to keep the position nonnegative. The process $\{X(t) : 0 \leq t \leq t_0\}$ of the player's position is an Itô process with reflection at zero, and its infinitesimal parameters μ and σ are chosen by the player at each instant of time from a control set depending on the current position. The probability of reaching the goal a by the time t_0 before exhausting all the fuel is maximized if the player can choose the parameters so that σ and μ/σ^2 are simultaneously maximized, at least when these maxima are sufficiently regular.

As an application of this control problem, a new comparison theorem for Itô processes with reflection is derived.

Key words. stochastic control, gambling, local time

AMS(MOS) subject classifications. 93E20, 60G40

1. Introduction. In this paper, we address a continuous-time stochastic control problem with finite fuel and a finite-time horizon. The same problem with an infinite-time horizon has been studied in [2]. This problem is inspired by a discrete-time problem in which a player with fortune x , a positive integer, seeks to reach a goal a , which is a larger positive integer. The random process of the player's subsequent fortunes depends upon control parameters, and, if the process reaches zero, the player can re-enter the game by paying a fee. The object is to maximize the probability of reaching the goal before arriving at zero without any funds left with which to pay the fee. In the continuous-time problem, we consider the payment of fees as the expenditure of fuel.

Let the player's position at time t be given by a stochastic process $X(t)$ with the state space $[0, a]$ and $X(0) = x$. The player's goal is to reach a ($0 \leq x < a$) before time t_0 , and he is given a finite amount of fuel, say $y > 0$. The fuel is spent in the cheapest possible way to keep the process $X(t)$ nonnegative. To do this, the process X is assumed to be reflecting at the origin, and its local time at the origin measures the expenditure of fuel. The dynamics are such that X is an Itô process with the control parameters $(\mu(\cdot), \sigma(\cdot))$ and is reflecting at the origin; see §2. The amount of fuel spent up to time t is measured by $L^X(t)$, the local time for X at the origin by time t . The problem is to choose the controls $(\mu(t), \sigma(t))$ from a given control set $C(X(t))$ so as to maximize the probability of reaching a goal a within a finite time t_0 before $L^X(\cdot)$ exceeds a level $y > 0$, i.e., to maximize $P[L^X(T_a^X) \leq y, T_a^X \leq t_0 | X(0) = x]$, where T_a^X is the first time $X(\cdot)$ reaches a . The control sets $\{C(z) : 0 \leq z \leq a\}$ are available to the player prior to the game.

When there is no finite time constraint, this problem is solved in [2]. Under some reasonable hypotheses on the family $\{C(z) : 0 \leq z \leq a\}$ in [2], an optimal choice for control parameters $\mu(\cdot)$ and $\sigma(\cdot)$ was given, and, for this optimal choice, the corresponding process $X(\cdot)$ was a diffusion with instantaneous reflection at zero. Furthermore, the optimal probability (i.e., the value function for the control problem) was explicitly computed, and the optimality of the process was established using a general verification lemma. This is in contrast with general stochastic control problems, where we can rarely compute the value function or explicitly determine the

* Received by the editors April 30, 1990; accepted for publication (in revised form) July 17, 1991.

† Department of Mathematics, Iowa State University, Ames, Iowa 50011.

optimal strategy.

In this paper, the solution becomes more difficult due to the finite time constraint $T_a^X \leq t_0$.

The methods and the proofs of this paper are different from and independent of those of the case of infinite-time horizon [2]. *Here we can guess the optimal strategy, but we cannot compute the value function explicitly.* If the player is away from the origin, he should use the controls to maximize the probability of reaching the goal a before hitting the origin. An optimal strategy for this problem (over a finite-time horizon) was given by Sudderth and Weerasinghe [19]. This observation led to our guess of the optimal strategy in this problem. In the process of verifying the optimality of our guess, we are led to a second-order partial differential equation (PDE) with mixed boundary data. The existence of a solution to such an equation can be proved using a result recently obtained by Lieberman [13]. The proof is given in the Appendix.

In the next section, we describe the results as well as the technical difficulties. The proofs are given in §3. A new comparison theorem for Itô processes is derived in §4.

There are a number of articles on finite-fuel problems available in the literature. (See, for example, [3], [8]–[11].) In these articles, it is usually assumed that μ and σ are known constants, and the player controls the use of fuel. In this paper, the player controls the parameters μ and σ . The two previous papers by Athreya and Weerasinghe deal with reflecting Itô processes over an infinite-time horizon [1], [2]. The nonreflecting case has been studied by Pestien and Sudderth [17], and Heath et al. [7], and by Sudderth and Weerasinghe [19] with a finite-time-horizon constraint. In the special case where $y = 0$, our problem reduces to that considered in [19]. In a recent work [20], Sudderth and Weerasinghe have considered a related problem for reflecting Itô processes with jumps (over an infinite-time horizon). In this article, the method of spending fuel to keep the process nonnegative is similar to that in a paper by Jacka [9].

2. Statement of the problem and results. Let $\{X(t) : t \geq 0\}$ be an Itô process with reflection at zero, initial position x , and parameters $\mu(\cdot)$ and $\sigma(\cdot)$. That is, $X(\cdot)$ is given by

$$\begin{aligned} X(t) &= A(t) + L^X(t), \\ A(t) &= x + \int_0^t \mu(s) ds + \int_0^t \sigma(s) dW(s), \\ L^X(t) &= -\min \left\{ \inf_{0 \leq u \leq t} A(u), 0 \right\}, \end{aligned} \tag{2.1}$$

where $W(\cdot)$ is a standard Brownian motion on some probability space (Ω, \mathcal{F}, P) adapted to a filtration such that $\mathcal{F}_t \subset \mathcal{F}$, and, for each $t \geq 0$, \mathcal{F}_t is independent of the Brownian increments $\{W(t+s) - W(t) : s \geq 0\}$ and contains all the P -null sets; $\mu(t, w)$ and $\sigma(t, w)$ are progressively measurable and satisfy

$$\int_0^t (|\mu(s)| + \sigma^2(s)) ds < \infty \quad \text{a.s. for each } t > 0. \tag{2.2}$$

The decomposition $X = A + L^X$ in (2.1) is known as the Skorohod decomposition [16]. It is easily verified that $L^X(\cdot)$ is nondecreasing and increases only when X is at zero, i.e.,

$$\int_0^t I(X(s) > 0) dL^X(s) = 0 \quad \forall t \geq 0.$$

The quantity $L^X(t)$ represents the amount of fuel used by the player (whose position is governed by the process X) up to time t to stay in the nonnegative half-line. This is the “cheapest” or “minimal” way to spend fuel to keep the process $X(t)$ nonnegative.

To make this more precise, suppose that there is another nondecreasing process $K(t)$ adapted to the same filtration such that $K(0) = 0$ and $A(t) + K(t) \geq 0$ for all t . Then it is easy to see that $K(t) \geq L^X(t)$.

Let $y > 0$ be the amount of fuel available to the player. The player's aim is to reach a fixed goal $a > x$ before exhausting the fuel $y > 0$ and within a finite-time horizon $t_0 > 0$.

Associated with every $z \in [0, a]$ is a control set $C(z)$, which is a nonempty subset of $R \times R^+$. The player is required to choose the value of (μ, σ) from $C(z)$ whenever the current position is z . More precisely, we assume that, for each t ,

$$(2.3) \quad (\mu(t), \sigma(t)) \in C(X(t)) \quad \text{whenever } 0 \leq X(t) \leq a.$$

Consider the stopping times T_a^X and τ_y^X , shown below:

$$(2.4) \quad \begin{aligned} T_a^X &= \inf\{t \geq 0 : X(t) \geq a\} \\ &= +\infty \text{ if the above set is empty} \end{aligned}$$

and

$$(2.5) \quad \begin{aligned} \tau_y^X &= \inf\{t \geq 0 : L^X(t) \geq y\} \\ &= +\infty \text{ if the above set is empty.} \end{aligned}$$

The player quits the game at the time instant $\min\{T_a^X, \tau_y^X, t_0\}$. In the event that $\{T_a^X \leq \min\{\tau_y^X, t_0\}\}$, the player wins the game; otherwise, he loses.

So the problem is the following. *Given the initial position $x \geq 0$, initial fuel amount $y > 0$, and the time limit $t_0 > 0$, the player would like to choose the processes $\mu(\cdot)$ and $\sigma(\cdot)$ so as to maximize the probability $K^X(x, y, t_0) = P_x[T_a^X \leq \tau_y^X \wedge t_0]$. (Here $a \wedge b$ denotes the quantity $\min\{a, b\}$).*

To know the amount of fuel available at time t , we introduce the fuel process by

$$(2.6) \quad Y(t) = y - L^X(t) \quad \text{for } t \geq 0.$$

Let

$$(2.7) \quad \rho(x) = \sup \left\{ \frac{\mu}{\sigma^2} : (\mu, \sigma) \in C(x) \right\}, \quad 0 \leq x \leq a,$$

where $C(x)$ is the control set available at x .

Now we make the following assumptions on the family of control sets $\{C(x) : 0 \leq x \leq a\}$.

Assumption 1. The function $\rho(\cdot)$ defined by (2.7) is continuous on $[0, a]$ and can be written in the form

$$(2.8) \quad \rho(x) = \mu_0(x)/\sigma_0^2(x), \quad 0 \leq x \leq a,$$

where $\mu_0(\cdot)$ and $\sigma_0(\cdot)$ are bounded continuous functions on $[0, a]$, $\inf_{[0, a]} \sigma_0(x) > 0$, and $(\mu_0(x), \sigma_0(x)) \in C(x)$ for every $x \in [0, a]$.

Assumption 2. $\sigma_0(\cdot)$ satisfies

$$(2.9) \quad \sigma_0(x) = \sup\{\sigma : \text{there exists } \mu \in R, \text{ such that } (\mu, \sigma) \in C(x)\}.$$

Assumption 3. For any process X satisfying (2.1) and with τ_y^X defined by (2.5),

$$\inf_{0 \leq s \leq u} \left(\int_{\tau_y^X}^{\tau_y^X + s} \mu(r) dr + \int_{\tau_y^X}^{\tau_y^X + s} \sigma(r) dW(r) \right) < 0$$

for all $u \in (0, t_0]$ with probability one.

Remark 1. Assumption 3 eliminates the use of the controls $\mu \equiv 0$ and $\sigma \equiv 0$ simultaneously in $[\tau_y^X, \tau_y^X + s]$ for some $s > 0$. Furthermore, if $\sigma(s)$ remains zero in $[\tau_y^X, \tau_y^X + s]$ for some $s > 0$, then, together with Assumption 1, it implies that $\mu(\cdot)$ is strictly negative in that same interval.

Assumption 3 is equivalent to $L(u + \tau_y) - L(\tau_y) > 0$ for all $u \in (0, t_0]$, with probability one. If X is a diffusion, then $L(\cdot)$ is the local time for X at the origin, and, since $X(\tau_y) = 0$, $L(\cdot + \tau_y) - L(\tau_y)$ gives the local time for the process $X(\cdot + \tau_y)$, therefore Assumption 3 holds automatically! If X is an Itô process, then we can verify Assumption 3 if there exists $u > 0$ such that $\sigma^2(s) > \epsilon_0 > 0$ for all $s \in [\tau_y, \tau_y + u]$. This can be easily done by using Girsanov's formula [12, p. 190].

Remark 2. Suppose that

$$\begin{aligned} \mu_0(x) &= \sup\{\mu : (\mu, \sigma) \in C(x) \text{ for some } \sigma\} \quad \text{and} \\ \sigma_0(x) &= \sup\{\sigma : (\mu, \sigma) \in C(x) \text{ for some } \mu\} \quad \text{for } 0 \leq x \leq a. \end{aligned}$$

Furthermore, assume that $\mu_0(\cdot)$ and $\sigma_0(\cdot)$ are continuous functions on $[0, a]$,

$$\sup_{[0, a]} \mu_0(x) \leq 0, \quad \inf_{[0, a]} \sigma_0(x) > 0, \quad \text{and} \quad (\mu_0(x), \sigma_0(x)) \in C(x) \quad \text{for every } x \in [0, a];$$

then Assumptions 1 and 2 are satisfied, and this example corresponds to an “unfavorable” or “subfair” game due to the negative drift (see [17]).

Let $\sum(x, y)$ be the collection of all processes $X = \{X(t) : t \geq 0\}$ given by (2.1) and satisfying Assumptions 1–3. These are the processes available to a player with initial position $x \geq 0$ and initial fuel supply $y > 0$.

The value function for this problem is defined by

$$(2.10) \quad V(x, y, t_0) = \sup_{X \in \sum(x, y)} P_x[T_a^X \leq \tau_y^X \wedge t_0].$$

Let $Z = \{Z(t) : t \geq 0\}$ be a reflecting diffusion process with instantaneous reflection at zero, satisfying

$$(2.11) \quad Z(t) = x + \int_0^t \mu_0(Z(s)) ds + \int_0^t \sigma_0(Z(s)) dW(s) + L^Z(t)$$

prior to reaching the goal $a > 0$, where $\{W(t) : t \geq 0\}$ is a Brownian motion on some probability space and $L^Z(t)$ is the local time of the $Z(t)$ process. So $L^Z(t)$ is a nondecreasing continuous process satisfying

$$\int_0^t I_{[Z(s) > 0]} dL^Z(s) = 0 \quad \text{for all } t \geq 0.$$

For existence and uniqueness of the solution to (2.11), see [18]; for the one-dimensional case, see [4].

Define

$$(2.12) \quad Q(x, y, t) = P_x[T_a^Z \leq \tau_y^Z \wedge t],$$

where T_a^Z and τ_y^Z are as in (2.4), (2.5) with the process X replaced by the process Z .

The following theorem is our main result.

THEOREM 1. *If Assumptions 1–3 hold, then $Q(x, y, t_0) = V(x, y, t_0)$ for all $x \in [0, a], y > 0$ and $t_0 \geq 0$. Therefore the process Z described in (2.11) is optimal.*

The proof of this result is somewhat long and involved, so we outline the major steps here.

Step 1. For each $X \in \sum(x, y)$, define its fuel process Y as in (2.6) and then define a process Y^ϵ by

$$Y^\epsilon(t) = Y(t) + \epsilon \int_0^t \sigma(s) dB(s),$$

where $B(\cdot)$ is a Brownian motion independent of the filtration $\{\mathcal{F}_t\}$. Let

$$\begin{aligned} \tau_y^\epsilon &= \inf\{t : t \geq 0, Y^\epsilon(t) = 0\}, \\ K_\epsilon^X(x, y, t) &= P(T_a^X \leq \tau_y^\epsilon \wedge t \mid X(0) = x), \\ K^X(x, y, t) &= K_0^X(x, y, t) = P[T_a^X \leq \tau_y \wedge t \mid X(0) = x]. \end{aligned}$$

In Lemma 3, we show that, for every $X \in \sum(x, y)$,

$$\lim_{\epsilon \rightarrow 0} K_\epsilon^X(x, y, t) = K^X(x, y, t).$$

Step 2. Let $\mu_n(\cdot)$ and $\sigma_n(\cdot)$ be functions on $[0, a]$ that are four times continuously differentiable and that satisfy

$$\frac{\mu_n(x)}{\sigma_n^2(x)} \geq \rho(x), \quad \sigma_n^2(x) \geq \sigma_0^2(x).$$

Let Z_n be the reflecting diffusion corresponding to (μ_n, σ_n) . We show (Lemma 2) that

$$K_\epsilon^{Z_n}(x, y, t) \geq K_\epsilon^X(x, y, t) \quad \forall X \in \sum(x, y).$$

Step 3. Choose μ_n, σ_n as in Step 2, which satisfies $\mu_n(\cdot) \rightarrow \mu_0(\cdot)$ and $\sigma_n(\cdot) \rightarrow \sigma_0(\cdot)$ uniformly on $[0, a]$. Then Z_n converges weakly to Z , and we show (in Lemma 4) that

$$\overline{\lim}_n K^{Z_n}(x, y, t) \leq K^Z(x, y, t) \equiv Q(x, y, t).$$

Combining Steps 1 and 2 and letting $\epsilon \rightarrow 0$, we obtain that $K^{Z_n}(x, y, t) \geq K^X(x, y, t)$ for all $X \in \sum(x, y)$, and then, letting n tend to infinity, we obtain that

$$\underline{\lim}_n K^{Z_n}(x, y, t) \geq K^X(x, y, t).$$

Finally, Step 3 yields that

$$\forall X \in \sum(x, y), Q(x, y, t) \geq \overline{\lim}_n K^{Z_n}(x, y, t) \geq \underline{\lim}_n K^{Z_n}(x, y, t) \geq K^X(x, y, t)$$

which is the assertion of Theorem 1.

It is Step 2 that is most crucial. Here we use a verification lemma and a recent result from the theory of (PDEs) that requires uniform ellipticity and smoothness of the coefficients in (3.7) and (3.8), below. It is precisely for this reason that we introduce the process Y^ϵ of Step 1. The other two steps are somewhat standard. Assumption 3 is crucial in Step 1, and the weak convergence of Z_n to Z is crucial in Step 3.

It should be noted that, with $\epsilon = 0$, (3.7) is a degenerate parabolic PDE, and the existence of a solution to such an equation with mixed boundary condition as in (3.8) is not available in the literature. The existence of a solution to (3.7) and (3.8) with $\epsilon > 0$ and smooth μ and σ needs the recent result of Lieberman [13] and is outlined in the Appendix.

In §4 we apply Theorem 1 to obtain a comparison theorem for reflecting Itô processes. This is related to a result of Hajek [6].

3. Approximate processes and a verification lemma. Throughout this section, we keep the constants a and t_0 fixed. We formulate the problem in three dimensions with state space

$$F = \{(x, y, t) : 0 \leq x \leq a, 0 \leq y, 0 \leq t \leq t_0\}.$$

Given a process $X \in \sum(x, y)$ satisfying (2.1), we can introduce the fuel process Y defined by (2.6).

Now, for each $\epsilon > 0$, we introduce a process Y^ϵ as follows:

$$(3.1) \quad Y^\epsilon(t) = y + \epsilon \int_0^t \sigma(s) dB(s) - L^X(t) \quad \text{for } 0 \leq t \leq T_a^X,$$

where the process $\sigma(s)$ is as in (2.1) and $\{B(t) : t \geq 0\}$ is a Brownian motion independent of the filtration $\{\mathcal{F}_t\}$. So we enlarge the filtration to include $\{B(t) : t \geq 0\}$.

Define

$$(3.2) \quad \begin{aligned} \tau_y^{Y^\epsilon} &= \inf\{t \geq 0 : Y^\epsilon(t) = 0\} \\ &= \infty \quad \text{if the above set is empty.} \end{aligned}$$

Recall T_a^X defined by (2.4). Whenever there can be no ambiguity, we write T_a and τ_y^ϵ , instead of T_a^X and $\tau_y^{Y^\epsilon}$, respectively, without identifying the processes X and Y^ϵ .

Now define the ϵ -approximate value function

$$(3.3) \quad V^\epsilon(x, y, t_0) = \sup_{X \in \sum(x, y)} P_x[T_a \leq \tau_y^\epsilon \wedge t_0]$$

LEMMA 1 (Verification lemma). *Let $G : F \rightarrow R$. Assume that G is continuous on F and has continuous second derivatives in F^0 , the interior of F . Let any $X \in \sum(x, y)$ satisfy (2.1), and let the corresponding Y^ϵ be defined by (3.1). Assume that*

(a)

$$\left(\frac{1}{2} \sigma^2(s) \left(\frac{\partial^2 G}{\partial x^2} + \epsilon^2 \frac{\partial^2 G}{\partial y^2} \right) + \mu(s) \frac{\partial G}{\partial x} - \frac{\partial G}{\partial s} \right) (X(s), Y^\epsilon(s), t_0 - s) \leq 0$$

for $0 < s < t_0 \wedge \tau_y^\epsilon$ with probability one;

(b)

$$\left(\frac{\partial G}{\partial x} - \frac{\partial G}{\partial y} \right) (0, Y^\epsilon(s), t_0 - s) \leq 0$$

for $0 < s < t_0 \wedge \tau_y^\epsilon$ with probability one;

(c) $G(x, y, s) \geq 0$ on F and $G(a, y, s) \geq 1$ for $0 \leq y, 0 \leq s \leq t_0 \wedge \tau_y^\epsilon$.
Then $G(x, y, t_0) \geq V^\epsilon(x, y, t_0)$.

In the following remark, we outline the idea of the proof.

Remark. $\{G(X(t), Y^\epsilon(t), t_0 - t) \mid 0 \leq t \leq t_0 \wedge \tau_y^\epsilon\}$ is a supermartingale from the conditions (a) and (b) of Lemma 1. Hence its expected value at the random time $t = t_0 \wedge T_a \wedge \tau_y^\epsilon$ is less than or equal to $G(x, y, t_0)$. The boundary conditions given in (c) implies that this expected value is greater than or equal to the probability $P_x[T_a \leq \tau_y^\epsilon \wedge t_0]$.

Proof. Consider that $G(X(t), Y^\epsilon(t), t_0 - t)$ for $0 \leq t \leq t_0 \wedge \tau_y^\epsilon$. Define the sequence (λ_n) of stopping times by

$$\begin{aligned} \lambda_n &= \inf\{t \geq 0 : |Y^\epsilon(t)| \geq n\} \\ &= +\infty \quad \text{otherwise.} \end{aligned}$$

Now, applying Itô's formula for $0 \leq t \leq t_0 \wedge \tau_y^\epsilon$, we obtain that

$$\begin{aligned} (3.4) \quad & G(X(t \wedge \lambda_n), Y^\epsilon(t \wedge \lambda_n), t_0 - (t \wedge \lambda_n)) \\ &= G(x, y, t_0) + \int_0^{t \wedge \lambda_n} \frac{\partial G}{\partial x}(X(s), Y^\epsilon(s), t_0 - s) \sigma(s) dW(s) \\ &+ \epsilon \int_0^{t \wedge \lambda_n} \frac{\partial G}{\partial y}(X(s), Y^\epsilon(s), t_0 - s) \sigma(s) dB(s) \\ &+ \int_0^{t \wedge \lambda_n} \left[\frac{1}{2} \sigma^2(s) \left(\frac{\partial^2 G}{\partial x^2} + \epsilon^2 \frac{\partial^2 G}{\partial y^2} \right) + \mu(s) \frac{\partial G}{\partial x} - \frac{\partial G}{\partial s} \right] (X(s), Y^\epsilon(s), t_0 - s) ds \\ &+ \int_0^{t \wedge \lambda_n} \left(\frac{\partial G}{\partial x} - \frac{\partial G}{\partial y} \right) (0, Y^\epsilon(s), t_0 - s) dL^X(s) \end{aligned}$$

By conditions (a) and (b), the last two terms in the right-hand side are less than or equal to zero. Furthermore, λ_n increases to $+\infty$, almost surely as $n \rightarrow +\infty$. Now replace t by $t_0 \wedge T_a \wedge \tau_y^\epsilon$, and, taking the expected values, we obtain that

$$\begin{aligned} E[G(X(t_0 \wedge T_a \wedge \tau_y^\epsilon \wedge \lambda_n), Y^\epsilon(t_0 \wedge T_a \wedge \tau_y^\epsilon \wedge \lambda_n), t_0 - (t_0 \wedge T_a \wedge \tau_y^\epsilon \wedge \lambda_n))] \\ \leq G(x, y, t_0) \end{aligned}$$

Now, letting $n \rightarrow +\infty$,

$$E[G(X(t_0 \wedge T_a \wedge \tau_y^\epsilon), Y^\epsilon(t_0 \wedge T_a \wedge \tau_y^\epsilon), t_0 - (t_0 \wedge T_a \wedge \tau_y^\epsilon))] \leq G(x, y, t_0).$$

Finally, using condition (c), we obtain that

$$P_x[T_a \leq \tau_y^\epsilon \wedge t_0] \leq \int_{[T_a \leq \tau_y^\epsilon \wedge t_0]} G(a, Y^\epsilon(T_a), t_0 - T_a) dP \leq G(x, y, t_0).$$

So it follows that $G(x, y, t_0) \geq V^\epsilon(x, y, t_0)$. \square

Our next step involves comparing the probabilities due to the approximate processes. Consider the functions $\mu_0(\cdot)$ and $\sigma_0(\cdot)$ in Assumptions 1 and 2 and the corresponding process Z defined by (2.11). Now we define the ϵ -approximate fuel process

$$(3.5) \quad Y^\epsilon(t) = y + \epsilon \int_0^t \sigma_0(Z(s)) dB(s) - L^Z(t) \quad \text{for } 0 \leq t \leq T_a^Z,$$

as in (3.1). Again, $B(\cdot)$ and $W(\cdot)$ are independent Brownian motions. Now T_a is defined by (2.4) for the process Z , and τ_y^ϵ is defined by (3.2).

The reason for defining the ϵ -approximate fuel process is to have uniform nondegeneracy of the partial differential operator involved in (3.7). This helps to guarantee a solution to (3.7) and (3.8).

Now define

$$(3.6) \quad Q^\epsilon(x, y, t) = P_x[T_a \leq \tau_y^\epsilon \wedge t] \quad \text{associated with the process } Z.$$

LEMMA 2. Assume that $\mu_0(\cdot)$ and $\sigma_0(\cdot)$ have continuous second derivatives on $[0, a]$; then, for each $t_0 > 0$,

$$Q^\epsilon(x, y, t_0) = V^\epsilon(x, y, t_0), \quad \text{where } V^\epsilon(x, y, t_0) \text{ is given by (3.3).}$$

Proof. Let $U_\epsilon(x, y, t)$ be the solution to

$$(3.7) \quad \frac{1}{2} \sigma_0^2(x) \left(\frac{\partial^2 U_\epsilon}{\partial x^2} + \epsilon^2 \frac{\partial^2 U_\epsilon}{\partial y^2} \right) + \mu_0(x) \frac{\partial U_\epsilon}{\partial x} = \frac{\partial U_\epsilon}{\partial t}$$

with the boundary conditions

$$(3.8) \quad \begin{aligned} U_\epsilon(x, y, 0) &= 0 & \text{for } 0 \leq x < a, y \geq 0, \\ U_\epsilon(x, 0, t) &= 0 & \text{for } x \geq 0, t \geq 0, \\ U_\epsilon(a, y, t) &= 1 & \text{for } y \geq 0, t \geq 0, \\ \left(\frac{\partial U_\epsilon}{\partial x} - \frac{\partial U_\epsilon}{\partial y} \right) (0, y, t) &= 0 & \text{for } y > 0, t \geq 0. \end{aligned}$$

Finding a solution to this system under C^2 -assumptions on $\mu_0(\cdot)$ and $\sigma_0(\cdot)$ is an easy extension of the work of Lieberman [13]–[15]. A proof of this and the fact that

$$(3.9) \quad U_\epsilon(x, y, t) = Q^\epsilon(x, y, t)$$

is presented in the Appendix. From (3.6) and (3.9), it is clear that $0 \leq U_\epsilon \leq 1$ and

$$(3.10) \quad \frac{\partial U_\epsilon}{\partial x} \geq 0 \quad \text{and} \quad \frac{\partial U_\epsilon}{\partial t} \geq 0 \quad \text{on } F^0.$$

It remains to show that $U_\epsilon(x, y, t_0) \geq V^\epsilon(x, y, t_0)$.

Now we would like to apply the verification lemma to the function U_ϵ . However, U_ϵ is not continuous on the edges of $\{(a, y, t) : y \geq 0, t \geq 0\}$. To avoid this difficulty, we define the function $W_n(x, y, t) = U_\epsilon(x, y + 1/n, t + 1/n)$ for $n = 1, 2, 3, \dots$.

Clearly, W_n is decreasing to U_ϵ , is continuous on F , and satisfies (3.7), together with the conditions

$$(3.11) \quad \begin{aligned} W_n(x, y, t) &\geq 0 && \text{on } F, \\ W_n(a, y, t) &= 1 && \text{for } y \geq 0, t \geq 0, \\ \left(\frac{\partial W_n}{\partial x} - \frac{\partial W_n}{\partial y} \right) (0, y, t) &= 0 && \text{for } y > 0, t \geq 0. \end{aligned}$$

Also, from (3.10), it follows that

$$(3.12) \quad \frac{\partial W_n}{\partial x} \geq 0 \quad \text{and} \quad \frac{\partial W_n}{\partial t} \geq 0 \quad \text{on } F^0.$$

Now take $X \in \Sigma(x, y)$ as in (2.1) and consider the corresponding $Y^\epsilon(\cdot)$ as in (3.1). To verify condition (a) of Lemma 1, consider that

$$\begin{aligned} &\left[\frac{1}{2} \sigma^2(s) \left(\frac{\partial^2 W_n}{\partial x^2} + \epsilon^2 \frac{\partial^2 W_n}{\partial y^2} \right) + \mu(s) \frac{\partial W_n}{\partial x} - \frac{\partial W_n}{\partial s} \right] (X(s), Y^\epsilon(s), t_0 - s) \\ &= \frac{1}{2} \sigma^2(s) \left[\frac{\partial^2 W_n}{\partial x^2} + \epsilon^2 \frac{\partial^2 W_n}{\partial y^2} + \frac{2\mu(s)}{\sigma^2(s)} \frac{\partial W_n}{\partial x} - \frac{2}{\sigma^2(s)} \frac{\partial W_n}{\partial s} \right] (X(s), Y^\epsilon(s), t_0 - s) \\ &\leq \frac{1}{2} \sigma^2(s) \left[\frac{\partial^2 W_n}{\partial x^2} + \epsilon^2 \frac{\partial^2 W_n}{\partial y^2} \right. \\ &\quad \left. + 2 \frac{\mu_0(X(s))}{\sigma_0^2(X(s))} \frac{\partial W_n}{\partial x} - \frac{2}{\sigma_0^2(X(s))} \frac{\partial W_n}{\partial s} \right] (X(s), Y^\epsilon(s), t_0 - s) \\ &= 0. \end{aligned}$$

The inequality is by (2.9), (2.10), and (3.12), and the final equality follows from the fact that W_n satisfy (3.7).

Conditions (b) and (c) of Lemma 1 clearly follow from (3.11).

Hence we conclude that $W_n(x, y, t_0) \geq V^\epsilon(x, y, t_0)$; by letting $n \rightarrow +\infty$, it follows that $U_\epsilon(x, y, t_0) \geq V^\epsilon(x, y, t_0)$; this, together with (3.9), proves that $Q^\epsilon(x, y, t_0) = V^\epsilon(x, y, t_0)$. \square

LEMMA 3. Let $X \in \Sigma(x, y)$ satisfy (2.1) and Assumptions 1–3, τ_y^X be as in (2.5), and τ_y^ϵ as in (3.2). Then $\lim_{\epsilon \rightarrow 0} P_x[T_a^X \leq \tau_y^\epsilon \wedge t_0] = P_x[T_a^X \leq \tau_y^X \wedge t_0]$ for all $t_0 \geq 0$.

Proof. Let X be as in (2.1) and keep $t_0 > 0$ fixed. With Y and Y^ϵ defined as in (2.6) and (3.1), respectively, we have that

$$(3.13) \quad \sup_{[0, t_0]} |Y(t \wedge T_a^X) - Y^\epsilon(t \wedge T_a^X)| \leq \epsilon \sup_{[0, t_0]} |h(t \wedge T_a^X)|,$$

where

$$h(t \wedge T_a^X) = \int_0^{t \wedge T_a^X} \sigma(s) dB(s).$$

Note that, as $\epsilon \rightarrow 0$, the right-hand side of (3.13) approaches zero. Also, the $Y(\cdot)$ process is decreasing. With Assumption 3, τ_y^X is a point of strict decrease for the process $Y(\cdot)$. Hence, with the aid of (3.13), we can easily see that $\tau_y^\epsilon \rightarrow \tau_y^X$ with probability one. Also, note that $T_a^X \neq \tau_y^X$ since $X(\tau_y^X) = 0$.

Therefore $\lim_{\epsilon \rightarrow 0} P_x[T_z^X \leq \tau_y^\epsilon \wedge t_0] = P_x[T_a^X \leq \tau_y^X \wedge t_0]$ for all $t_0 > 0$. \square

Now we can prove Theorem 1 under additional smoothness assumptions on μ_0 and σ_0 . Then we use weak convergence to prove it for the general case when μ_0 and σ_0 are continuous on $[0, a]$.

Proof of Theorem 1. (With smoothness assumptions). Here, in addition to the assumptions stated in Theorem 1, we assume that the functions $\mu_0(\cdot)$ and $\sigma_0(\cdot)$ have continuous second derivatives on $[0, a]$. Now let Z be the process given by (2.11), and $Q(x, y, t)$ be given by (2.12). Furthermore, let $Q^\epsilon(x, y, t)$ be as in (3.6). By Lemma 3, it is clear that $Q^\epsilon(x, y, t)$ converges to $Q(x, y, t)$. Now take $X \in \Sigma(x, y)$; define $P_x[T_a^X \leq \tau_y^\epsilon \wedge t_0]$ and $P_x(T_a^X \leq \tau_y^X \wedge t_0]$ as before. By Lemma 2, $P_x[T_a^X \leq \tau_y^\epsilon \wedge t_0] \leq Q^\epsilon(x, y, t_0)$. Now letting $\epsilon \rightarrow 0$ in both sides and using Lemma 3, we conclude that $P_x[T_a^X \leq \tau_y^X \wedge t_0] \leq Q(x, y, t_0)$. Hence, as $Z \in \Sigma(x, y)$, it follows that $V(x, y, t_0) = Q(x, y, t_0)$, as desired. So, with the smoothness assumptions, the proof of Theorem 1 is complete. \square

To remove our additional assumption on the smoothness of $\mu_0(\cdot)$ and $\sigma_0(\cdot)$, we need a result on weak convergence.

Let $\mu_0(\cdot)$ and $\sigma_0(\cdot)$ be continuous on $[0, a]$ and extend them to $[0, \infty)$ as bounded continuous functions (simply let $\mu_0(x) = \mu_0(a)$ and $\sigma_0(x) = \sigma_0(a)$ for $x \geq a$).

Now approximate μ_0 and σ_0 by bounded C^∞ functions μ_n and σ_n on $[0, a]$, and hence on $[0, \infty)$, such that

$$(3.14) \quad \begin{aligned} \sigma_n &\geq \sigma_0, \quad \rho_n = \mu_n / \sigma_n^2 \geq \mu_0 / \sigma_0^2 = \rho \\ \sup_{0 \leq x} (|\mu_n(x) - \mu_0(x)| + |\sigma_n(x) - \sigma_0(x)|) &\leq \frac{1}{n}, \end{aligned}$$

and the functions $|\mu_n|$, $|\mu_0|$, $|\sigma_0|$, $|\sigma_n|$ for $n = 1, 2, \dots$ are all bounded by a constant $M > 0$.

For the given continuous $\mu_0(\cdot)$ and $\sigma_0(\cdot)$, we define

$$(3.15) \quad \begin{aligned} X(t) &= x + \int_0^t \mu_0(X(s)) ds + \int_0^t \sigma_0(X(s)) dW(s) + L^X(t), \\ Y(t) &= y - L^X(t) \end{aligned}$$

on some probability space (Ω, \mathcal{F}, P) with respect to some Brownian motion $W(t)$ on this space. For this process X , we define T_a and τ_y as before. Now, for each $n = 1, 2, \dots$, since μ_n and σ_n are smooth we can define

$$(3.16) \quad \begin{aligned} X_n(t) &= x + \int_0^t \mu_n(X_n(s)) ds + \int_0^t \sigma_n(X_n(s)) dW(s) + L^{X_n}(t), \\ Y_n(t) &= y - L^{X_n}(t). \end{aligned}$$

The processes X_n and Y_n are all defined on the same probability space (Ω, \mathcal{F}, P) with respect to the same Brownian motion $W(\cdot)$. We denote $T_a^{X_n}$ by T_a^n , and $\tau_y^{X_n}$ by τ_y^n . We restrict our time interval to $[0, t_0]$. Furthermore, we introduce the sup norm $\|\cdot\|_{[0, t_0]}$ on $C([0, t_0] \rightarrow \mathbf{R})$ by $\|f\|_{[0, t_0]} = \sup_{0 \leq t \leq t_0} |f(t)|$.

LEMMA 4. Define the processes (X, Y) and (X_n, Y_n) by (3.15) and (3.16), respectively. Then

(a) There exists a subsequence n_p such that (X_{n_p}, Y_{n_p}) converges weakly to (X, Y) , and

(b) $\limsup_{n_p \rightarrow +\infty} P[T_a^{n_p} \leq \tau_y^{n_p} \wedge t_0 | X_{n_p}(0) = x] \leq P[T_a \leq \tau_y \wedge t_0 | X(0) = x]$.

Proof. Part (a) follows from Theorem 3.2 of Lions and Sznitman [16] and the uniqueness of solutions to the submartingale problem of Stroock and Varadhan [18]. To derive part (b), let ν_0 be the probability measure induced by (X, Y) on the space of continuous functions $C([0, t_0] \rightarrow R^2)$, and ν_n be the corresponding probability measure for (X_n, Y_n) .

Now, for a given continuous $f : [0, t_0] \rightarrow R$, define $T_a^f = \inf\{t > 0 : f(t) \geq a\}$. Let

$$A = \{(f, g) \in C([0, t_0] \rightarrow R^2) : \|f\|_{[0, t_0]} \geq a, g \text{ is decreasing, } g \geq 0 \text{ on } [0, T_a^f]\}.$$

Clearly, A is closed with respect to the sup norm. Let $(f_n, g_n) \in A$ converge to (f, g) ; then f and g are continuous, and g is decreasing. Furthermore, $\|f_n\|_{[0, t_0]} \geq a$ implies that $\|f\|_{[0, t_0]} \geq a$ and $T_a^f \leq \liminf_{n \rightarrow \infty} T_a^{f_n}$. Hence $g_n \geq 0$ on $[0, T_a^f]$, which, in turn, implies that $g \geq 0$ on $[0, T_a^f]$. Therefore A is closed.

Next we show that $\nu_n(A) = P[T_a^n \leq \tau_y^n \wedge t_0 | X_n(0) = x]$ and $\nu_0(A) = P[T_a \leq \tau_y \wedge t_0 | X(0) = x]$.

Since the proof is essentially the same for all n , we show the following for ν_0 : Consider the paths of the diffusion (X, Y) corresponding to parameters $\mu_0(\cdot)$ and $\sigma_0(\cdot)$,

$$\begin{aligned} w \in [T_a \leq \tau_y \wedge t_0] &\iff \|X(t, w)\|_{[0, t_0]} \geq a, & \tau_y(w) \geq T_a(w) \\ &\iff \|X(t, w)\|_{[0, t_0]} \geq a, & Y(t, w) > 0 \text{ on } [0, T_a(w)] \\ &\iff \|X(t, w)\|_{[0, t_0]} \geq a, & Y(t, w) \geq 0 \text{ on } [0, T_a(w)]. \end{aligned}$$

The last implication holds because, if $Y(r, w) = 0$ for some $r < T_a(w)$, then (since $X(\cdot)$ is a diffusion satisfying (3.15)), we have that $Y(s) < 0$ for $r < s < T_a(w)$. Also, see the explanation after Assumption 3.

Hence $w \in [T_a \leq \tau_y \wedge t_0] \iff (X(t, w), Y(t, w)) \in A$. Now, by part (a) of Lemma 3, ν_{n_p} converges to ν_0 as the subsequence n_p tends to infinity. Since A is a closed set, $\limsup_{n_p \rightarrow \infty} \nu_{n_p}(A) \leq \nu_0(A)$. This proves the lemma. \square

Proof of Theorem 1. (With only the continuity assumptions on μ_0 and σ_0). Consider the subsequence n_p given in Lemma 3. For each n_p , $\mu_{n_p}/\sigma_{n_p}^2 \geq \mu_0/\sigma_0^2$ and $\sigma_{n_p} \geq \sigma_0$, and μ_{n_p}, σ_{n_p} are bounded C^∞ functions. Hence, from the proof of Theorem 1 with smoothness assumptions on coefficients, it follows that $V(x, y, t_0) \leq P[T_a^{n_p} \leq \tau_y^{n_p} \wedge t_0] = \nu_{n_p}(A)$. Hence $V(x, y, t_0) \leq \limsup_{n_p \rightarrow +\infty} \nu_{n_p}(A) \leq \nu_0(A) = Q(x, y, t_0)$.

Furthermore, it is clear that $Q(x, y, t_0) \leq V(x, y, t_0)$ since X , given by (2.15), satisfies $X \in \Sigma(x, y)$. This completes the proof. \square

4. A comparison theorem. Consider the processes X and Z satisfying (2.1) and (2.11), respectively. Hajek [4, Thm. 2] showed that, if σ_0 is a constant and if

$$X(0) \leq Z(0), \quad \mu(t) \leq \mu_0(X(t)), \quad |\sigma(t)| \leq \sigma_0,$$

then $P[X_t \geq c] \leq 2 \cdot P[Z_t \geq c]$ for every $t \geq 0$ and $c \geq 0$.

Now we formulate a related comparison theorem for reflecting Itô processes. Let

$$(4.1) \quad X(t) = x + \int_0^t \sigma(s) dW(s) + \int_0^t \mu(s) ds + L^X(t), x \geq 0$$

as in (2.1), and

$$(4.2) \quad Z(t) = z + \int_0^t \sigma_0(Z(s)) dB(s) + \int_0^t \mu_0(Z(s)) ds + L^Z(t), z \geq 0,$$

where $L^Z(\cdot)$ is the local time of Z at zero as in (2.11). Note that the Brownian motions $W(\cdot)$ and $B(\cdot)$ in (4.1) and (4.2) may not be related. Furthermore, the processes $X(\cdot)$ and $Z(\cdot)$ may be defined on different probability spaces.

THEOREM 2. *Let X and Z be defined by (3.1) and (3.2), respectively. Assume that $a > 0$ and that*

- (i) $X(0) \leq Z(0)$;
- (ii)

$$\frac{\mu(s)}{\sigma^2(s)} \leq \frac{\mu_0(X(s))}{\sigma_0^2(X(s))} \quad \text{and} \quad \sigma^2(s) \leq \sigma_0^2(X(s))$$

for all s whenever $0 \leq X(s) \leq a$;

(iii) $\mu_0(\cdot)$ and $\sigma_0(\cdot)$ are continuous, and $\sigma_0^2(\cdot)$ is bounded below by a positive constant $\epsilon > 0$;

(iv) Assumption 3 remains true for X for a sequence $\{y_n\}$ increasing to $+\infty$.

Then, for each $t > 0$, $P[\sup_{[0,t]} X(s) \geq a] \leq P[\sup_{[0,t]} Z(s) \geq a]$.

Proof. First, consider the case where $X(0) = Z(0)$. Then, for each n , $P[T_a^x \leq \tau_{y_n}^x \wedge t] \leq P[T_a^z \leq \tau_{y_n}^z \wedge t]$ from Theorem 1.1. Letting y_n increase to $+\infty$, we have that $P[T_a^x \leq t] \leq P[T_a^z \leq t]$; so $P[\sup_{[0,t]} X(s) \geq a] \leq P[\sup_{[0,t]} Z(s) \geq a]$. If $X(0) = x < Z(0) = z$, then consider an auxiliary diffusion process Z^1 satisfying (4.2) with the initial condition $Z^1(0) = x$. In this case where $P[\sup_{[0,t]} X(s) \geq a] \leq P[\sup_{[0,t]} Z^1(s) \geq a]$. From (3.6), (3.9), and (3.10), together with Lemma 3, it follows that

$$P[\sup_{[0,t]} Z^1(s) \geq a] \leq P[\sup_{[0,t]} Z(s) \geq a].$$

This completes the proof. \square

Appendix. Here we would like to sketch a proof for the solution of the PDE (3.7) with the boundary data (3.8). We assume that $\mu_0(\cdot)$ and $\sigma_0(\cdot)$ have continuous second derivatives. To give this solution, we follow the work of Lieberman [13]–[15], and our result is an easy extension of his work. Let $M > 1$ be an integer. Define a continuous decreasing function ϵ_M on $[0, M]$, so that

$$\epsilon_M(y) = 1 \quad \text{for} \quad 0 \leq y \leq M - 1.$$

ϵ_M is decreasing on $[M - 1, M]$, and $\epsilon_M(M) = 0$. Furthermore, let $\phi(\cdot)$ be a continuous increasing function on $[0, 1]$ such that $\phi(0) = 0$ and $\phi(1) = 1$.

Define the domain D_M by

$$(A.1) \quad D_M = \{(x, y, t) : 0 < x < a, 0 < y < M, t > 0\}.$$

Now let U_M be the solution to (3.7) on the domain D_M satisfying the continuous boundary data

$$(A.2) \quad \begin{aligned} U_M(a, y, t) &= \epsilon_M(y) \quad 0 \leq y \leq M, & t \geq 0; \\ U_M(x, M, t) &= 0 \quad 0 \leq x \leq 1 & t \geq 0; \\ U_M(x, 0, t) &= \phi(x) \quad 0 \leq x \leq 1, & t \geq 0; \\ U_M(x, y, 0) &= \phi(x)\epsilon_M(y) \quad 0 \leq x \leq 1, \quad 0 \leq y \leq M; \\ \left(\frac{\partial U_M}{\partial x} - \frac{\partial U_M}{\partial y} \right) (0, y, t) &= 0 \quad 0 < y < M, & t \geq 0. \end{aligned}$$

The existence of the solution U_M follows from [13]. So, as $M = 2, 3, \dots$, we obtain a sequence (U_M) of solutions. Let L be the operator

$$L = \frac{1}{2} \sigma_0^2(x) \left(\frac{\partial^2}{\partial x^2} + \epsilon^2 \frac{\partial^2}{\partial y^2} \right) + \mu_0(x) \frac{\partial}{\partial x} - \frac{\partial}{\partial t}.$$

Applying the maximum principle with the operator L (see Friedman [5, Chap. 2, Thms. 2, 14, pp. 38, 49]), we can easily obtain the following conclusions:

(i) $0 \leq U_M \leq 1$ on D_M , and

(ii) $U_M \leq U_{M+1}$ on D_M .

Hence, for each (x, y, t) such that $0 \leq x \leq a, 0 \leq y, t \geq 0$, $\{U_M(x, y, t)\}$ is an increasing sequence in M and is bounded above by 1. Let $U(x, y, t)$ be the limit. Define

$$(A.3) \quad D_\infty = \{(x, y, t) : 0 < x < a, 0 < y, 0 < t\}.$$

We would like to show that U satisfies (3.7) on D_∞ . Take any small ball $B \subseteq D_\infty$ such that $(x, y, t) \in B$; then $D_M \supseteq B$ for all $M > M_0$: Introduce H_2, H_0 , and $H_{2+\alpha}$ norms on B (for details, see [14]) for $0 < \alpha < 1$. Then

$$(A.4) \quad \|U_n - U_m\|_{H_2} \leq C \cdot \|U_n - U_m\|_{H_0}^{2/(2+\alpha)} \cdot \|U_n - U_m\|_{H_{2+\alpha}}^{\alpha/(2+\alpha)} \quad \text{for } n, m > M_0.$$

(For this, see (2.1a) and (2.1b) of [14].) Since $U_n \nearrow U$, however, $\|U_n - U_m\|_{H_0}$ approaches zero. Furthermore,

$$(A.5) \quad \|U_n - U_m\|_{H_{2+\alpha}} \leq K \cdot \|U_n - U_m\|_{H_0}$$

in [5, Chap. 3, Thm. 5, p. 64]. Hence (U_n) is a Cauchy sequence in H_2 , and, for each $n > M_0$, U_n satisfies (3.7). Therefore U also satisfies (3.7). So $0 \leq U \leq 1$ on D_∞ , and it solves (3.7) on D_∞ with the boundary data

$$\begin{aligned} U(x, 0, t) &= \phi(x), & 0 \leq x \leq a, t \geq 0, \\ U(x, y, 0) &= \phi(x), & 0 \leq x \leq a, y \geq 0, \\ U(a, y, t) &= 1, & y \geq 0, t \geq 0, \end{aligned}$$

and

$$\left(\frac{\partial U}{\partial x} - \frac{\partial U}{\partial y} \right) (0, y, t) = 0.$$

Now, instead of ϕ , we choose a sequence $\{\phi_n\}$ of smooth increasing functions so that $\phi_n(0) = 0, \phi_n(a) = 1$, and, for each $0 \leq x < a$, $\phi_n(x)$ is decreasing to zero as $n \rightarrow +\infty$; call this solution U_n . So $0 \leq U_n \leq 1$.

By Itô's formula applied to $(Z(t), Y^\epsilon(t))$, where Z is defined by (2.11) and Y^ϵ is defined by (2.5), we obtain that

$$E[U_n(Z(s), Y^\epsilon(s), t-s)] = U_n(x, y, t),$$

since $\{U_n(Z(s), Y^\epsilon(s), t-s) : 0 \leq s \leq t\}$ is a martingale. So we derive $U_n(x, y, t) = E[U_n(Z(T_a \wedge \tau_y^\epsilon \wedge t), Y^\epsilon(T_a \wedge \tau_y^\epsilon \wedge t), t - (T_a \wedge \tau_y^\epsilon \wedge t))]$. This shows that

$$U_n(x, y, t) = P_x[T_a \leq \tau_y^\epsilon \wedge t] + \int_{[\tau_y^\epsilon < T_a \wedge t]} \phi_n(Z_{\tau_y^\epsilon}) dP + \int_{[t < T_a \wedge \tau_y^\epsilon]} \phi_n(Z_t) dP.$$

Since ϕ_n is decreasing to zero, clearly, $U_n(x, y, t)$ is decreasing to $Q^\epsilon(x, y, t) \equiv P_x[T_a \leq \tau_y^\epsilon \wedge t]$, as defined in (3.6). Now, following inequalities (5.4), (5.5), and the same argument given there, $Q^\epsilon(x, y, t)$ satisfies (3.7) on D_∞ with the boundary data given by (3.8). \square

Acknowledgments. The author thanks Professor K. B. Athreya for many valuable discussions and for his keen interest in this work. Thanks are also due to Professor G. M. Lieberman for telling me about his work related to the content of this paper, and to the referee, whose comments have helped improve the readability of this paper.

REFERENCES

- [1] K. B. ATHREYA AND A. WEERASINGHE *Exponentiality of the local time at hitting times for reflecting diffusions and an application*, in Probability, Statistics and Mathematics: Papers in Honor of Samuel Karlin, T. W. Anderson et al., eds., Academic Press, New York, 1989.
- [2] ———, *Reflecting Itô processes in a stochastic control problem*, Math. Oper. Res., to appear.
- [3] V. E. BENEŠ, L. A. SHEPP, AND H. S. WITSSENHAUSEN, *Some solvable stochastic control problems*, Stochastics, 4 (1980), pp. 39–83.
- [4] M. CHALEYAT-MAUREL AND N. EL-KAROUI, *Un problème de réflexion et ses applications au temps local et aux équations différentielles stochastiques sur \mathbb{R} , cas continu*, Asterisque, 52 (1978), pp. 117–144.
- [5] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Krieger, Malabar, Florida, 1983.
- [6] B. HAJEK, *Mean comparison of diffusions*, Z. Wahrsch. View. Gebiete, 68 (1985), pp. 315–329.
- [7] D. HEATH, S. OREY, V. PESTIEN, AND W. SUDDERTH, *Minimizing or maximizing the expected time to reach zero*, SIAM J. Control Optim., 25 (1987), pp. 195–205.
- [8] D. C. HEATH AND W. D. SUDDERTH, *Continuous-time gambling problems*, Adv. Appl. Probab., 6 (1974), pp. 651–665.
- [9] S. D. JACKA, *A finite fuel stochastic control problem*, Stochastics, 10 (1983), pp. 103–113.
- [10] I. KARATZAS, *Probabilistic aspects of finite-fuel stochastic control*, Proc. Nat. Acad. Sci. USA, 82 (1985), pp. 5579–5581.
- [11] I. KARATZAS AND S. E. SHREVE, *Equivalent models for finite-fuel stochastic control*, Stochastics, 18 (1986), pp. 245–276.
- [12] ———, *Brownian Motion and Stochastic Calculus*, Graduate Texts in Math, Springer-Verlag, Berlin, New York, 1988.
- [13] G. LIEBERMAN, *Mixed boundary value problems for elliptic and parabolic differential equations of second-order*, J. Math. Anal. Appl., 113 (1986), pp. 422–440.
- [14] ———, *Intermediate Schauder theory for second-order parabolic equations I. Estimates*, J. Differential Equations, 63 (1986), pp. 1–31.
- [15] ———, *Intermediate Schauder theory for second-order parabolic equations II, Existence, uniqueness and regularity*, J. Differential Equations, 63 (1986), pp. 32–57.
- [16] P. L. LIONS AND A. S. SZNITMAN, *Stochastic differential equations with reflecting boundary conditions*, Comm. Pure Ap. Math., 37 (1984), pp. 511–537.
- [17] V. C. PESTIEN AND W. D. SUDDERTH, *Continuous-time red and black: How to control a diffusion to a goal*, Math. Oper. Res., 10 (1983), pp. 599–611.
- [18] D. W. STROOCK, AND S. R. S. VARADHAN, *Diffusion processes with boundary conditions*, Comm. Pure Ap. Math., 24 (1971), pp. 147–225.
- [19] W. D. SUDDERTH AND A. WEERASINGHE, *Controlling a process to a goal in finite time*, Math. Oper. Res., 14 (1989), pp. 400–409.
- [20] ———, *Using fuel to control a process to a goal*, Stochastics Stochastics Rep., 34 (1989), pp. 169–186.

SENSITIVITY ANALYSIS OF PARAMETRIZED PROGRAMS UNDER CONE CONSTRAINTS*

A. SHAPIRO[†] AND J. F. BONNANS[‡]

Abstract. In this paper local behavior of optimal solutions of parametrized optimization problems is investigated with cone constraints in Banach spaces. Under second-order sufficient optimality conditions Lipschitzian stability of the corresponding ε -optimal solutions is established. Also shown is how the considered parametric program can be approximated by using second-order expansions of the involved functions.

Key words. nonlinear optimization, parametric programming, stability and sensitivity analysis, second-order optimality conditions, Lipschitz continuity

AMS(MOS) subject classifications. 49K40, 90C31

1. Introduction. In this paper we study local behavior of ε -optimal solutions of the parametric optimization problem

$$(\mathcal{P}_t) \quad \min_{x \in X} f(x, t) \quad \text{subject to } x \in \Phi(t),$$

with $f : X \times \mathbb{R}_+ \rightarrow \mathbb{R}$, depending on the parameter $t \geq 0$. Here X is a real Banach space, $t \in \mathbb{R}_+$, and it will be supposed throughout the paper that the feasible set $\Phi(t)$ is defined by cone constraints. That is,

$$(1.1) \quad \Phi(t) = \{x \in X : g(x, t) \in K\},$$

where $g : X \times \mathbb{R}_+ \rightarrow Y$, Y is a Banach space and K is a closed convex cone in Y . With the program (\mathcal{P}_t) is associated the optimal value function

$$(1.2) \quad \varphi(t) := \inf\{f(x, t) : x \in \Phi(t)\}$$

and an ε -optimal ($\varepsilon > 0$) solution $\bar{x}(t)$ satisfying the conditions $\bar{x}(t) \in \Phi(t)$ and

$$(1.3) \quad f(\bar{x}(t), t) \leq \varphi(t) + \varepsilon.$$

Program (\mathcal{P}_t) can be considered in a general context of parametric optimization. That is, suppose that the objective function and the constraint mapping depend on a parameter $u \in U$, where U is a vector space. Then we can study directional behavior of the corresponding optimal solutions. That is, we restrict ourselves to the investigation of the one-parameter family of optimization problems by considering perturbations tu_0 , $t \in \mathbb{R}_+$, in a given direction $u_0 \in U$.

The main result of this paper is given in Theorem 1, where we show that under certain second-order sufficient conditions and for $\varepsilon = \varepsilon(t)$ tending to zero sufficiently quickly as $t \rightarrow 0^+$, the corresponding ε -optimal solutions $\bar{x}(t)$ are upper Lipschitz continuous at $t = 0$. This extends some recent results in sensitivity analysis of nonlinear programs ([4]–[6], [8], [15]) to the considered infinite-dimensional case and cone

* Received by the editors December 5, 1990; accepted for publication (in revised form) July 29, 1991.

[†] School of Industrial and Systems Engineering, Georgia Institute of Technology, Georgia 30332-0205.

[‡] Institut National de Recherche en Informatique et en Automatique, Domaine de Voluceau, BP 105, 78153 Rocquencourt, France.

constraints. In §3 we show how the problem (\mathcal{P}_t) can be approximated by a simpler one that involves second-order expansions of $f(x, t)$ and $g(x, t)$. This approximation may prove to be useful in the calculation of directional derivatives of $\bar{x}(t)$ and requires further investigation.

We assume that $f(x, t)$ and $g(x, t)$ are *twice continuously Fréchet differentiable*, jointly in x and t , and denote by $D_x f(x, t)$, $D_x g(x, t)$, $D_{tt}^2 g(x, t)$, etc., the corresponding partial derivatives. In particular, $D_{xx}^2 g(x, 0)$ belongs to the space $\mathcal{L}(X, \mathcal{L}(X, Y))$ of bounded linear operators from X to the space $\mathcal{L}(X, Y)$, equipped with the corresponding operator norm, and we denote $[D_{xx}^2 g(x, 0)y]y$ by $D_{xx}^2 g(x, 0)(y, y)$. We suppose throughout that the null program (\mathcal{P}_0) has a *unique* optimal solution x_0 and that x_0 is a *regular* point of $g(x) = g(x, 0)$, with respect to the cone K , in the sense of Robinson [13]. That is

$$(1.4) \quad 0 \in \text{int} \{g(x_0) + Dg(x_0)X - K\}.$$

The following notation and terminology will be used in the paper. By $\text{cl}\{S\}$ we denote the topological closure of a set $S \subset X$. For a convex set C and $x \in C$ we denote

$$R(x, C) := \bigcup_{t \geq 0} t(C - x)$$

the radial cone, and $T(x, C) := \text{cl}\{R(x, C)\}$ the tangent cone to C at x . $B(x; r)$ denotes the ball $\{y \in X : \|y - x\| \leq r\}$ and $B_X = B(0; 1)$ denotes the unit ball in X . For $x \in X$ and $\xi \in X^*$ we use the notation $\langle \xi, x \rangle$ or $\langle x, \xi \rangle$ for the value $\xi(x)$ of the linear functional ξ at x . If K is a cone in X or in X^* , then its (positive) dual cone K^+ is given by

$$K^+ := \{y : \langle y, x \rangle \geq 0 \text{ for all } x \in K\},$$

and its polar (negative dual) cone is $K^- := -K^+$. By $\partial f(x)$ we denote the subdifferential of a convex function $f(x)$. For a point $x \in X$ and a set $S \subset X$ we denote by $\text{dist}(x, S)$ the distance from x to S .

2. Lipschitz stability of optimal solutions. In this section we study Lipschitz continuity of ε -optimal solutions of (\mathcal{P}_t) . Note that we assumed existence of the optimal solution for the null (unperturbed) program (\mathcal{P}_0) only. Of course, for $\varepsilon > 0$, an ε -optimal solution $\bar{x}(t)$ always exists provided that the corresponding feasible set $\Phi(t)$ is nonempty and $\varphi(t) > -\infty$. Sometimes we write $f(x)$, $g(x)$ and Φ for $f(x, 0)$, $g(x, 0)$ and $\Phi(0)$, respectively.

Under the regularity assumption (1.4) the set

$$(2.1) \quad \Lambda_0 := \{\lambda \in K^+ : Df(x_0) = \lambda \circ Dg(x_0), \langle \lambda, g(x_0) \rangle = 0\}$$

of Lagrange multipliers of the program (\mathcal{P}_0) at the optimal solution point x_0 is nonempty (first-order necessary conditions, [11], [14]) and bounded (e.g. [18, Thm. 4.1]). Consequently Λ_0 is a convex and weakly* compact subset of Y^* . Consider the Lagrangian function

$$L(x, \lambda, t) := f(x, t) - \langle \lambda, g(x, t) \rangle$$

of the program (\mathcal{P}_t) and the set

$$(2.2) \quad \Lambda_1 := \arg \max \{D_t L(x_0, \lambda, 0) : \lambda \in \Lambda_0\}.$$

Notice that the set Λ_1 is nonempty because of the weak* compactness of Λ_0 . We need the following regularity assumption.

Assumption A. For some $\lambda_0 \in \Lambda_1$ the tangent cone $T(\lambda_0, \Lambda_0)$ is representable in the form

$$(2.3) \quad T(\lambda_0, \Lambda_0) = \{\lambda \in T(\lambda_0, K^+) : \lambda \circ Dg(x_0) = 0, \quad \langle \lambda, g(x_0) \rangle = 0\}.$$

It follows from the definition (2.1) of the set Λ_0 that the radial cone $R(\lambda_0, \Lambda_0)$ can be written in the form

$$(2.4) \quad R(\lambda_0, \Lambda_0) = \{\lambda \in R(\lambda_0, K^+) : \lambda \circ Dg(x_0) = 0, \quad \langle \lambda, g(x_0) \rangle = 0\}.$$

Therefore, Assumption A holds if and only if the topological closure of the cone given in the right-hand side of (2.4) coincides with the cone given in the right-hand side of (2.3). In particular, Assumption A holds in the following cases:

(i) The cone K^+ satisfies at λ_0 the polyhedral property

$$(2.5) \quad T(\lambda_0, K^+) = R(\lambda_0, K^+).$$

(ii) The point x_0 is regular with respect to the cone $K_0 := K(\lambda_0)$, where

$$(2.6) \quad K(\lambda) := \{y \in K : \langle \lambda, y \rangle = 0\}.$$

Condition (2.5) is satisfied in the situations where $\lambda_0 = 0$ or when the feasible set $\Phi(t)$ is defined by equality constraints and a *finite* number of inequality constraints, i.e., $Y = Y_1 \times \mathbb{R}^n$ and $K = \{0\} \times \mathbb{R}_+^m$, where Y_1 is a Banach space and 0 is the zero vector of Y_1 . Condition (ii) was considered in [16]. It implies that the set $\Lambda_0 = \{\lambda_0\}$ is a singleton ([16, Lemma 4.3]) and hence $T(\lambda_0, \Lambda_0) = \{0\}$. Also, condition (ii) is equivalent to (see [11, Lemma 2.3])

$$(2.7) \quad Dg(x_0)X - K_0 + [g(x_0)] = Y,$$

where $[g(x_0)]$ denotes the linear space generated by vector $g(x_0)$. The polar cone of the cone $Dg(x_0)X - K_0 + [g(x_0)]$ is the intersection of the polar cones of $Dg(x_0)X$, $-K_0$ and $[g(x_0)]$. Since

$$-K_0^- = \text{cl}(K^+ + [\lambda_0]) = T(\lambda_0, K^+),$$

it follows that the polar cone of $Dg(x_0)X - K_0 + [g(x_0)]$ coincides with the cone given in the right-hand side of (2.3). By (2.7) this polar cone is $\{0\}$ and hence, indeed, condition (ii) implies Assumption A. Later we will give an example where there exists a unique Lagrange multiplier but Assumption A does not hold. We note, however, that if (2.5) holds and the cone $Dg(x_0)X - K_0 + [g(x_0)]$ is closed, then condition (ii) is equivalent to the uniqueness of the multiplier (see Lemma 4.3 in [16]).

LEMMA 1. *Suppose that x_0 is a regular point of $g(x)$ with respect to K , that Assumption A holds, and that the cone $Dg(x_0)X - K_0 + [g(x_0)]$ is closed. Then there exist positive numbers κ and η such that*

$$(2.8) \quad \varphi(t) - \varphi(0) \leq t \max_{\lambda \in \Lambda_0} D_t L(x_0, \lambda, 0) + \kappa t^2$$

for all $t \in [0, \eta]$.

Proof. Since λ_0 maximizes $D_t L(x_0, \lambda, 0)$ over Λ_0 , it follows by the corresponding first-order necessary conditions that

$$D_\lambda[D_t L(x_0, \lambda, 0)] = -D_t g(x_0, 0) \in N(\lambda_0, \Lambda_0),$$

where $N(\lambda_0, \Lambda_0)$ is the normal cone to Λ_0 at λ_0 . The normal cone $N(\lambda_0, \Lambda_0)$ is polar of the cone $T(\lambda_0, \Lambda_0)$ and, because of Assumption A, is given by the topological closure of the cone $Dg(x_0)X - K_0 + [g(x_0)]$. Since it is assumed that the last cone is closed we obtain that

$$-D_t g(x_0, 0) \in Dg(x_0)X - K_0 + [g(x_0)].$$

It follows that for any $t > 0$ there exists $\bar{y} \in X$, $k \in K_0$ and $\alpha \in \mathbb{R}$ such that

$$-tD_t g(x_0, 0) = tDg(x_0)\bar{y} - k + \alpha g(x_0).$$

Moreover, since $g(x_0) \in K_0$ we can always take $\alpha \geq 0$. Therefore, for sufficiently small $t > 0$ we can take $\alpha \in [0, 1]$. Then replacing k by $k + (1 - \alpha)g(x_0) \in K_0$ we obtain

$$-tD_t g(x_0, 0) = tDg(x_0)\bar{y} - k + g(x_0).$$

It follows that

$$(2.9) \quad g(x_0) + tD_x g(x_0, 0)\bar{y} + tD_t g(x_0, 0) \in K,$$

and

$$(2.10) \quad \langle \lambda_0, D_x g(x_0, 0)\bar{y} + D_t g(x_0, 0) \rangle = 0.$$

Let us note that since K is convex and $g(x_0) \in K$, if (2.9) holds for some $t = t_0 > 0$ and \bar{y} , then it holds for all $t \in [0, t_0]$ and the same \bar{y} . Therefore, we can choose \bar{y} independently of t for t sufficiently small. Now

$$(2.11) \quad g(x_0 + t\bar{y}, t) = g(x_0) + tD_x g(x_0, 0)\bar{y} + tD_t g(x_0, 0) + O(t^2).$$

Consequently, by the Robinson–Ursescu stability theorem ([13], [17]), it follows from (2.9) and (2.11) that there exists $\tilde{y}(t)$ such that $x_0 + \tilde{y}(t) \in \Phi(t)$ and $\|t\bar{y} - \tilde{y}(t)\|$ is of order $O(t^2)$. Then

$$\begin{aligned} \varphi(t) &\leq f(x_0 + \tilde{y}(t), t) = f(x_0) + D_x f(x_0, 0)\tilde{y}(t) + tD_t f(x_0, 0) + O(t^2), \\ &= f(x_0) + tD_x f(x_0, 0)\bar{y} + tD_t f(x_0, 0) + O(t^2). \end{aligned}$$

Together with (2.10) this implies

$$\begin{aligned} \varphi(t) - \varphi(0) &\leq tD_x f(x_0, 0)\bar{y} - t\langle \lambda_0, D_x g(x_0, 0)\bar{y} \rangle \\ &\quad + tD_t f(x_0, 0) - t\langle \lambda_0, D_t g(x_0, 0) \rangle + O(t^2) \\ &= t(Df(x_0) - \lambda_0 \circ Dg(x_0))\bar{y} + tD_t L(x_0, \lambda_0, 0) + O(t^2). \end{aligned}$$

Since λ_0 maximizes $D_t L(x_0, \lambda, 0)$ over Λ_0 , the inequality (2.8) follows. \square

Remark 1. Condition (2.9) holds for some $t > 0$ if and only if

$$(2.12) \quad D_x g(x_0, 0)\bar{y} + D_t g(x_0, 0) \in K + [g(x_0)].$$

Note that

$$K + [g(x_0)] = R(g(x_0), K).$$

Now consider the following linearization of the program (\mathcal{P}_t) :

$$(\mathcal{L}_i) \quad \min_{y \in X} \langle Df(x_0), y \rangle + D_t f(x_0, 0) \quad \text{subject to } D_x g(x_0, 0)y + D_t g(x_0, 0) \in M_i,$$

$i = 1, 2$, where $M_1 = R(g(x_0), K)$ and $M_2 = T(g(x_0), K)$. Program (\mathcal{L}_2) differs from program (\mathcal{L}_1) only in that its feasible set is the topological closure of the feasible set of program (\mathcal{L}_1) . Clearly, \bar{y} solves program (\mathcal{L}_1) if and only if it satisfies condition (2.12) and solves program (\mathcal{L}_2) . Therefore, a feasible point \bar{y} solves (\mathcal{L}_1) if and only if there exists $\lambda \in [T(g(x_0), K)]^+$ such that

$$Df(x_0) = \lambda \circ Dg(x_0) \text{ and } \langle \lambda, Dg(x_0)\bar{y} + D_t g(x_0, 0) \rangle = 0.$$

(Note that regularity of the program (\mathcal{L}_2) follows from the regularity of the optimal solution x_0 of the null program (\mathcal{P}_0)). Since

$$[T(g(x_0), K)]^+ = K^+ \cap \text{Ker } g(x_0),$$

we obtain that \bar{y} solves program (\mathcal{L}_1) if and only if there exists $\lambda_0 \in \Lambda_0$ and $t > 0$ such that conditions (2.9) and (2.10) hold. It follows that under the assumptions of Lemma 1, program (\mathcal{L}_1) has a solution.

Remark 2. The dual of programs (\mathcal{L}_i) defined in Remark 1 is the problem of maximization of $D_t L(x_0, \lambda, 0)$ subject to $\lambda \in \Lambda_0$. By arguments similar to those of Lempio and Maurer ([9, pp. 142–143]), it is possible to show that under the assumption of regularity of x_0 , *alone*, there is no duality gap between programs (\mathcal{L}_i) and their dual program. Let us briefly outline those arguments.

Consider the optimal value function

$$\psi(v) = \inf \{ \langle Df(x_0), y \rangle : Dg(x_0)y + v \in T(g(x_0), K) \}.$$

This is a sublinear (convex and positively homogeneous) function and by the first-order optimality conditions $\psi(0) = 0$. Moreover, it follows from the generalized open mapping theorem [12] that $\psi(v)$ is bounded from above by a finite constant for all v in a neighborhood of zero. This implies that $\psi(v)$ is continuous (e.g. [7, Lemma 2.1]) and hence is a support function of a bounded set. We have that $\mu \in \partial\psi(0)$ if and only if $\psi(v) \geq \langle \mu, v \rangle$ for all $v \in Y$. By duality

$$\psi(v) \geq \max_{\lambda \in \Lambda_0} \langle -\lambda, v \rangle$$

and hence $-\Lambda_0 \subset \partial\psi(0)$. Also, by taking $y = 0$ in the definition of $\psi(v)$ we obtain that for $\mu \in \partial\psi(0)$,

$$\langle \mu, v \rangle \leq 0 \quad \text{for all } v \in T(g(x_0), K),$$

and hence

$$-\mu \in T(g(x_0), K)^+ = K^+ \cap \text{Ker } g(x_0).$$

Furthermore, for a given y taking $v = -Dg(x_0)y$, we obtain

$$\langle Df(x_0), y \rangle + \langle \mu, Dg(x_0)y \rangle \geq 0,$$

and hence

$$Df(x_0) + \mu \circ Dg(x_0) = 0.$$

Consequently, $-\mu \in \Lambda_0$ and hence $\partial\psi(0) = -\Lambda_0$. It follows that

$$\psi(v) = \max_{\lambda \in \Lambda_0} \langle -\lambda, v \rangle.$$

The result that there is no duality gap between program (\mathcal{L}_1) and its dual, implies the inequality (see [9, Thm. 3.1])

$$\varphi(t) - \varphi(0) \leq t \max_{\lambda \in \Lambda_0} D_t L(x_0, \lambda, 0) + o(t).$$

The above inequality is similar to the inequality (2.8) but with the term κt^2 replaced by $o(t)$. To derive the stronger inequality (2.8) we need existence of the optimal solution for (\mathcal{L}_1) , which is ensured by the assumptions of Lemma 1.

We employ the following second-order sufficient condition. For $\eta \geq 0$ consider the cone

$$(2.13) \quad C_\eta = \{y \in X : Dg(x_0)y \in K + [g(x_0)], \langle Df(x_0), y \rangle \leq \eta \|y\|\},$$

and the set Λ_1 defined in (2.2).

Assumption B (Second-order sufficient condition). There exists $\alpha > 0$ and $\eta > 0$ such that

$$(2.14) \quad \max_{\lambda \in \Lambda_1} \langle y, D_{xx}^2 L(x_0, \lambda, 0)y \rangle \geq \alpha \|y\|^2, \quad \text{for all } y \in C_\eta.$$

For $\eta = 0$ the corresponding cone C_0 is called the critical cone of the program (\mathcal{P}_0) . Note that by the first-order necessary conditions, $\langle Df(x_0), y \rangle$ is nonnegative for all y such that $Dg(x_0)y$ belongs to the radial cone $R(g(x_0), K)$ (e.g., [11, Thm. 3.1]). Therefore, the second-term inequality in the right-hand side definition (2.13) of C_0 can be replaced by the equation $\langle Df(x_0), y \rangle = 0$. The second-order sufficient condition of Assumption B is a natural extension of the (strong) second-order sufficient condition employed in [15, p. 635] for finite dimensional cases and a finite number of constraints.

Assumption B implies that for any positive number β less than $\alpha/2$, there exists a neighborhood W of x_0 such that

$$f(x) \geq f(x_0) + \beta \|x - x_0\|^2$$

for all $x \in \Phi \cap W$ (cf. [11, Thm. 5.6]). It follows then that if $\bar{x}(t)$ is an $\varepsilon(t)$ -optimal solution of (\mathcal{P}_t) , $\varepsilon(t) = O(t)$ and $\bar{x}(t) \in W$, then $\bar{x}(t)$ converges to x_0 at $t \rightarrow 0^+$ at least at a rate of $O(t^{1/2})$ ([1]–[3], [16]). In the following theorem we establish Lipschitzian rate of convergence of $\bar{x}(t)$ to x_0 .

THEOREM 1. *Let $\varepsilon(t) = O(t^2)$ and $\bar{x}(t)$ be an $\varepsilon(t)$ -optimal solution of (\mathcal{P}_t) converging to x_0 as $t \rightarrow 0^+$. Suppose that the assumptions of Lemma 1 and Assumption B hold. Then there exists a positive constant c such that*

$$(2.15) \quad \|\bar{x}(t) - x_0\| \leq ct$$

for all $t \geq 0$ sufficiently small.

Proof. Suppose that (2.15) is false. Then there are $t_n \rightarrow 0^+$, $x_n = \bar{x}(t_n)$ and $\tau_n = \|x_n - x_0\|$ such that

$$(2.16) \quad \lim_{n \rightarrow \infty} t_n / \tau_n = 0.$$

Let η be a positive constant specified in Assumption B and consider $y_n = \tau_n^{-1}(x_n - x_0)$. We have by (2.16) that

$$g(x_n, t_n) = g(x_0) + \tau_n Dg(x_0)y_n + o(\tau_n).$$

Since $g(x_n, t_n) \in K$ it follows then that

$$\text{dist}(Dg(x_0)y_n, K + [g(x_0)]) \rightarrow 0.$$

By the generalized open mapping theorem (or the Robinson–Ursescu stability theorem) this implies that $\text{dist}(y_n, S) \rightarrow 0$ where

$$S = \{y \in X : Dg(x_0)y \in K + [g(x_0)]\}.$$

Therefore, there exists $\bar{y}_n \in S$ such that $\|y_n - \bar{y}_n\|$ tends to zero as $n \rightarrow \infty$. Now by the definition of ε -optimality

$$\varphi(t_n) - \varphi(0) \geq f(x_n, t_n) - f(x_0, 0) - \varepsilon_n,$$

where $\varepsilon_n = \varepsilon(t_n)$. It follows then by (2.16) that

$$\varphi(t_n) - \varphi(0) \geq \tau_n \langle y_n, Df(x_0) \rangle + o(\tau_n).$$

Moreover, because of (2.8),

$$\limsup_{n \rightarrow \infty} \tau_n^{-1} [\varphi(t_n) - \varphi(0)] \leq 0.$$

It follows that

$$\langle \bar{y}_n, Df(x_0) \rangle \leq \eta,$$

for n large enough, and hence $\bar{y}_n \in C_\eta$.

Now since for every $\lambda \in \Lambda_0$,

$$\varphi(0) = L(x_0, \lambda, 0)$$

and

$$\varphi(t_n) \geq L(x_n, \lambda, t_n) - \varepsilon_n$$

we have that

$$\varphi(t_n) - \varphi(0) \geq \max_{\lambda \in \Lambda_0} \{L(x_n, \lambda, t_n) - L(x_0, \lambda, 0)\} - \varepsilon_n.$$

Furthermore,

$$\begin{aligned} L(x_n, \lambda, t_n) - L(x_0, \lambda, 0) &= t_n D_t L(x_0, \lambda, 0) + \frac{1}{2} \tau_n^2 \langle y_n, D_{xx}^2 L(x_n^*, \lambda, t_n^*) y_n \rangle \\ &\quad + t_n \tau_n \langle y_n, D_{xt}^2 L(x_n^*, \lambda, t_n^*) \rangle + \frac{1}{2} t_n^2 D_{tt}^2 L(x_n^*, \lambda, t_n^*), \end{aligned}$$

where (x_n^*, t_n^*) is a point on the segment joining $(x_0, 0)$ and (x_n, t_n) . It follows then by continuity of the second-order derivatives, boundedness of Λ_0 , and (2.16) that

$$\begin{aligned}\varphi(t_n) - \varphi(0) &\geq \max_{\lambda \in \Lambda_0} \{t_n D_t L(x_0, \lambda, 0) + \tfrac{1}{2} \tau_n^2 \langle y_n, D_{xx}^2 L(x_0, \lambda, 0) y_n \rangle\} + o(\tau_n^2), \\ &\geq t_n \max_{\lambda \in \Lambda_0} \{D_t L(x_0, \lambda, 0)\} + \tfrac{1}{2} \tau_n^2 \max_{\lambda \in \Lambda_1} \{\langle y_n, D_{xx}^2 L(x_0, \lambda, 0) y_n \rangle\} + o(\tau_n^2).\end{aligned}$$

Since $\|y_n - \bar{y}_n\|$ tends to zero and $\bar{y}_n \in C_\eta$ for n large enough we have by the second-order condition of Assumption B that

$$\max_{\lambda \in \Lambda_1} \langle y_n, D_{xx}^2 L(x_0, \lambda, 0) y_n \rangle \geq \tfrac{1}{2} \alpha,$$

and hence

$$\varphi(t_n) - \varphi(0) \geq t_n \max_{\lambda \in \Lambda_0} \{D_t L(x_0, \lambda, 0)\} + \tfrac{1}{4} \alpha \tau_n^2 + o(\tau_n^2).$$

The last inequality contradicts the result (2.8) of Lemma 1 and hence the proof is complete. \square

Lipschitzian stability of optimal solutions of parametrized programs was studied in a recent paper of Alt [2]. His result ([2, Thm. 3.4]) requires a regularity condition involving Lagrange multipliers that may be difficult to verify in situations where the set Λ_0 is not a singleton. This question has also been addressed by Malanowski [10] under different hypotheses.

As we mentioned earlier, if the point x_0 is regular with respect to the cone $K_0 = K(\lambda_0)$, given in (2.6), then $\Lambda_0 = \{\lambda_0\}$ and Assumption A holds. In this case, Lipschitzian stability of optimal solutions of (\mathcal{P}_t) implies Lipschitzian stability of the corresponding Lagrange multipliers ([16, Lemma 4.4]). It is interesting to note that regularity of x_0 with respect to the cone K , uniqueness of λ_0 , and Lipschitzian stability of the optimal solutions do not imply Lipschitzian stability of the corresponding Lagrange multipliers even in the finite dimensional case (hence does not imply Assumption A). We show this in the following example.

Example. Let $X = \mathbb{R}^2$, $Y = \mathbb{R}^3$ and consider

$$f(x) = x_1 + x_2 + x_1^2 + x_2^2,$$

$g(x) = Gx$, where G is the 3×2 matrix

$$G^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

and the cone

$$K = \{y \in \mathbb{R}^3; y_3^2 \geq y_1^2 + y_2^2, y_3 \geq 0\}.$$

Notice that $K^+ = K$. Then $x_0 = (0, 0)^t$ is the optimal solution of the problem

$$\text{minimize } f(x) \text{ subject to } g(x) \in K.$$

The corresponding first-order necessary conditions are

$$(\lambda_1, 0) + (0, \lambda_3) = (1, 1), \quad \lambda = (\lambda_1, \lambda_2, \lambda_3)^t \in K^+,$$

which give the unique Lagrange multiplier $\lambda_0 = (1, 0, 1)^t$. It can be easily verified that the point $x_0 = (0, 0)^t$ is a regular point of $g(x)$ with respect to the cone K but not with respect to the cone K_0 .

Consider the following perturbations of the objective function

$$f(x, t) = (1 - t)x_1 + (1 + t)x_2 + x_1^2 + x_2^2.$$

Then for all $t \in [0, 1]$, $x_0 = (0, 0)^t$ is the optimal solution with the corresponding set of Lagrange multipliers

$$\Lambda(t) = \{(1 - t, \lambda_2, 1 + t) : |\lambda_2| \leq 2t^{1/2}\}.$$

In particular, $\bar{\lambda}(t) = (1 - t, 2t^{1/2}, 1 + t)^t$ is a vector of the Lagrange multipliers with $\|\bar{\lambda}(t) - \lambda_0\| \geq t^{1/2}$.

3. Second-order expansion of (\mathcal{P}_t) . In this section we show how second-order expansions of $f(x, t)$ and $g(x, t)$ can be employed to approximate the program (\mathcal{P}_t) by a simpler one. Consider $a_0 = Df(x_0)$, $G_0 = Dg(x_0)$, $c_0 = g(x_0)$ and for $\eta \geq 0$ the set

$$(3.1) \quad \Lambda_\eta := \bigcup_{(a, G, c) \in \Omega_\eta} \{\lambda \in K^+ ; a = \lambda \circ G ; \langle \lambda, c \rangle = 0\},$$

where

$$\Omega_\eta := \{(a, G, c) ; \|a - a_0\| + \|G - G_0\| + \|c - c_0\| \leq \eta\}.$$

Clearly, Λ_η contains the set Λ_0 of Lagrange multipliers and for $\eta = 0$ both sets coincide. Note that under the assumption (1.4) of regularity of x_0 , the set Λ_η is bounded for sufficiently small $\eta > 0$. Indeed, consider $\lambda \in \Lambda_\eta$ and for a given $\varepsilon > 0$ let $y \in B_Y$ be such that $\langle \lambda, y \rangle \geq \|\lambda\| - \varepsilon$. Then by the generalized open mapping theorem, it follows from the regularity of x_0 that there are a positive constant α , $x \in B_X$, $|\gamma| \leq 1$ and $k \in K$ such that

$$y = \alpha(G_0x + \gamma c_0) - k$$

and $\alpha \leq \bar{\alpha}$, where $\bar{\alpha}$ does not depend on y .

Let $(a, G, c) \in \Omega_\eta$ be such that $a = \lambda \circ G$ and $\langle \lambda, c \rangle = 0$. We have then that

$$\begin{aligned} \langle \lambda, y \rangle &\leq \alpha \langle \lambda, G_0x + \gamma c_0 \rangle \\ &= \alpha \langle \lambda, Gx + \gamma c \rangle + \alpha \langle \lambda, (G_0 - G)x + \gamma(c_0 - c) \rangle, \\ &\leq \alpha \|a\| + \alpha \|\lambda\| (\|G_0 - G\| + \|c_0 - c\|), \\ &\leq \alpha \|a_0\| + \alpha \eta + \alpha \eta \|\lambda\|. \end{aligned}$$

It follows that for $\eta < \bar{\alpha}^{-1}$,

$$\|\lambda\| \leq (1 - \bar{\alpha}\eta)^{-1}(\bar{\alpha}\|a_0\| + \bar{\alpha}\eta + \varepsilon)$$

which shows that Λ_η is bounded.

We shall need the following strong form of second-order sufficient conditions.

Assumption C. There exist $\beta > 0$ and $\eta > 0$ such that for any $\lambda \in \Lambda_\eta$,

$$(3.2) \quad \langle y, D_{xx}^2 L(x_0, \lambda, 0)y \rangle \geq \beta \|y\|^2$$

for all $y \in X$.

Note that Assumption C implies the existence of a quadratic form on the space X that induces a norm equivalent to the original norm $\|\cdot\|$. Endowed with this new norm, X is a Hilbert space.

Now let us consider the program,

$$\begin{aligned} (Q_t) \quad & \min_{y \in X} \langle y, D_x f(x_0, t) \rangle + \frac{1}{2} \langle y, D^2 f(x_0) y \rangle \\ & \text{subject to } g(x_0, t) + D_x g(x_0, t) y + \frac{1}{2} D_{xx}^2 g(x_0, 0)(y, y) \in K. \end{aligned}$$

Note that under the assumptions of Theorem 1, it follows from the Lipschitz stability of $\bar{x}(t)$ that the optimal value function $\varphi(t)$ is differentiable at $t = 0$ (in the positive direction) and

$$\varphi'(0) = \max_{\lambda \in \Lambda_0} D_t L(x_0, \lambda, 0)$$

(cf. [9, Thm. 3.4]). A similar result holds for the program (Q_t) as well, hence the difference between the optimal value functions of programs (P_t) and (Q_t) is equal to $f(x_0, 0)$ plus a term of order $o(t)$.

THEOREM 2. *Suppose that the assumptions of Lemma 1 and Assumption C hold and that for all sufficiently small $t \geq 0$, program (Q_t) has an optimal solution $y^*(t)$. Let, for $\varepsilon(t) = o(t^2)$, $\bar{x}(t)$ be an $\varepsilon(t)$ -optimal solution of (P_t) converging to x_0 as $t \rightarrow 0^+$. Then*

$$(3.3) \quad \|\bar{x}(t) - x_0 - y^*(t)\| = o(t).$$

To prove Theorem 2 we use the following variational principle ([16, Lemma 2.2]). Let X be a normed space, S and T be subsets of X , $f, g : X \rightarrow \mathbb{R}$ be Lipschitz continuous and Gâteaux differentiable functions and consider optimization problems

$$(3.4) \quad \min_{x \in S} f(x)$$

and

$$(3.5) \quad \min_{x \in T} g(x).$$

Suppose that the program (3.4) has an optimal solution x_0 and that there exist an open, convex neighborhood W of x_0 and a constant $\gamma > 0$ such that

$$(3.6) \quad f(x) \geq f(x_0) + \gamma \|x - x_0\|^2$$

for all $x \in S \cap W$ (i.e., the cost satisfies a quadratic growth condition on the feasible set). Then for any ε -optimal solution $\bar{x} \in W$ of (3.5) we have

$$(3.7) \quad \|\bar{x} - x_0\| \leq \gamma^{-1} \kappa + \gamma^{-1/2} \varepsilon^{1/2} + 2\delta_1 + \gamma^{-1/2} (k_1 \delta_1 + k_2 \delta_2)^{1/2},$$

where k_1 and k_2 are Lipschitz constants of $f(x)$ and $g(x)$ in the neighborhood W , respectively,

$$\begin{aligned} \kappa &:= \sup\{\|Df(x) - Dg(x)\| : x \in W\}, \\ \delta_1 &:= \sup_{x \in T'} \text{dist}(x, S'), \\ \delta_2 &:= \text{dist}(x_0, T'), \end{aligned}$$

and $S' := S \cap W$, $T' := T \cap W$.

Proof of Theorem 2. We show that programs (\mathcal{P}_t) and (\mathcal{Q}_t) are sufficiently close to each other in terms of the upper bound (3.7). Consider the functions

$$\begin{aligned} f^*(y, t) &:= \langle y, D_x f(x_0, t) \rangle + \frac{1}{2} \langle y, D^2 f(x_0) y \rangle, \\ g^*(y, t) &:= g(x_0, t) + D_x g(x_0, t) y + \frac{1}{2} D_{xx}^2 g(x_0, 0)(y, y), \end{aligned}$$

the Lagrangian

$$L^*(y, \lambda, t) := f^*(y, t) - \langle \lambda, g^*(y, t) \rangle,$$

and the feasible set

$$\Psi(t) := \{y \in X : g^*(y, t) \in K\}$$

corresponding to the program (\mathcal{Q}_t) . We have that for any $\lambda \in \Lambda_0$ and $y \in \Psi(0)$,

$$f^*(y, 0) \geq L^*(y, \lambda, 0).$$

Since for any $\lambda \in \Lambda_0$,

$$L^*(y, \lambda, 0) = \frac{1}{2} \langle y, D_{xx}^2 L(x_0, \lambda, 0) y \rangle,$$

it follows then from Assumption C that

$$(3.8) \quad f^*(y, 0) \geq \frac{1}{2} \beta \|y\|^2$$

for all $y \in \Psi(0)$, and hence $y = 0$ is the optimal solution of (\mathcal{Q}_0) .

We show now that $y^*(t) \rightarrow 0$ as $t \rightarrow 0^+$. Since x_0 is regular we have by the Robinson–Ursescu stability theorem that there exists $v(t) \in \Psi(t)$ such that $v(t) \rightarrow 0$ as $t \rightarrow 0^+$. It follows that

$$f^*(y^*(t), t) \leq c(t),$$

where $c(t) = f^*(v(t), t)$ tends to zero as $t \rightarrow 0^+$. Moreover, we have that for any $\lambda \in \Lambda_0$,

$$\begin{aligned} f^*(y^*(t), t) &\geq L^*(y^*(t), \lambda, t) = L^*(y^*(t), \lambda, 0) + L^*(y^*(t), \lambda, t) - L^*(y^*(t), \lambda, 0) \\ &= \frac{1}{2} \langle y^*(t), D_{xx}^2 L(x_0, \lambda, 0) y^*(t) \rangle + \langle y^*(t), D_x f(x_0, t) - D_x f(x_0, 0) \rangle \\ &\quad - \langle \lambda, (D_x g(x_0, t) - D_x g(x_0, 0)) y^*(t) \rangle - \langle \lambda, g(x_0, t) - g(x_0, 0) \rangle, \\ &\geq \frac{1}{2} \beta \|y^*(t)\|^2 - a(t) \|y^*(t)\| - b(t), \end{aligned}$$

where

$$a(t) := \|D_x f(x_0, t) - D_x f(x_0, 0)\| + \|\lambda\| \|D_x g(x_0, t) - D_x g(x_0, 0)\|$$

and

$$b(t) := \langle \lambda, g(x_0, t) - g(x_0, 0) \rangle$$

tend (uniformly over λ) to zero at $t \rightarrow 0^+$. It follows that

$$\frac{1}{2} \beta \|y^*(t)\|^2 - a(t) \|y^*(t)\| - b(t) \leq c(t).$$

By solving this quadratic inequality with respect to $\|y^*(t)\|$, we obtain that $\|y^*(t)\| \rightarrow 0$ as $t \rightarrow 0^+$.

Now, by Theorem 1, we have that $\|\bar{x}(t) - x_0\|$ and $\|y^*(t)\|$ are of order $O(t)$. Therefore, there is $c > 0$ such that $\|\bar{x}(t) - x_0\| \leq ct$ and $\|y^*(t)\| \leq ct$ for all sufficiently small $t \geq 0$. By continuity of $D_x f(x, t)$ at $(x_0, 0)$ we have that $f(\cdot, t)$ and $f^*(\cdot, t)$ are Lipschitz continuous in neighbourhoods of x_0 and 0, respectively, with Lipschitz constants independent of t for all t small enough.

Let us estimate the constants

$$\delta_1(t) := \sup\{\text{dist}(x, \Phi(t)) : x \in (x_0 + \Psi(t)) \cap B(x_0; ct)\},$$

and

$$\delta_2(t) := \sup\{\text{dist}(y, \Psi(t)) : y \in (\Phi(t) - x_0) \cap B(0; ct)\}.$$

Since x_0 is regular we have by the Robinson–Ursescu stability theorem that for all (x, t) sufficiently close to $(x_0, 0)$ the distance $\text{dist}(x, \Phi(t))$ is of order $\text{dist}(g(x, t), K)$. If, in addition, $y = x - x_0 \in \Psi(t)$ and thus $g^*(y, t) \in K$, then

$$\text{dist}(g(x, t), K) \leq \|g(x, t) - g^*(y, t)\|.$$

Now by Taylor's theorem

$$g(x, t) - g^*(y, t) = \frac{1}{2} D_{xx}^2 g(x^*, t)(y, y) - \frac{1}{2} D_{xx}^2 g(x_0, 0)(y, y),$$

where x^* is a point on the segment joining x_0 and x . By continuity of $D_{xx}^2 g(x, t)$ this implies that if $x \in B(x_0; ct)$, then $\|g(x, t) - g^*(y, t)\|$ is of order $o(t^2)$. It follows then that $\delta_1(t) = o(t^2)$. By similar arguments $\delta_2(t) = o(t^2)$.

Furthermore, consider

$$\begin{aligned} \kappa(t) &:= \sup_{\|x - x_0\| \leq ct} \|D_x f(x, t) - D_x f^*(x - x_0, t)\|, \\ &= \sup_{\|x - x_0\| \leq ct} \|D_x f(x, t) - D_x f(x_0, t) - D^2 f(x_0, 0)(x - x_0)\|, \end{aligned}$$

which is of order $o(t)$ because of second-order Fréchet differentiability of $f(x, t)$. Now, for a given $t \geq 0$ small enough, consider $y^* = y^*(t)$. Because of the regularity of x_0 and since $y^*(t) \rightarrow 0$ as $t \rightarrow 0^+$, we have that there is a Lagrange multiplier $\lambda^* = \lambda^*(t)$ corresponding to the optimal solution $y^*(t)$ of the program (\mathcal{Q}_t) such that $\lambda^* \in \Lambda_\eta$ for sufficiently small t . It follows that for all $y \in \Psi(t)$,

$$(3.9) \quad f^*(y, t) - f^*(y^*, t) \geq L^*(y, \lambda^*, t) - L^*(y^*, \lambda^*, t).$$

By taking the second-order Taylor expansion of $L^*(\cdot, \lambda^*, t)$ at y^* and noting that because of the first-order optimality conditions $D_y L^*(y^*, \lambda^*, t) = 0$ and since the function is quadratic this expansion is exact, we obtain that the right-hand side of (3.9) is equal to $\frac{1}{2} \langle y - y^*, D_{xx}^2 L(x_0, \lambda^*, 0)(y - y^*) \rangle$. Consequently, Assumption C implies that

$$f^*(y, t) - f^*(y^*, t) \geq \frac{1}{2} \beta \|y - y^*\|^2$$

for all $y \in \Psi(t)$ and all t small enough. Now (3.3) follows from the inequality (3.7) applied to the programs (\mathcal{Q}_t) and (\mathcal{P}_t) . \square

Remark 3. In the case where

$$(3.10) \quad \lim_{\eta \rightarrow 0^+} \left\{ \sup_{\lambda \in \Lambda_\eta} \text{dist}(\lambda, \Lambda_0) \right\} = 0,$$

the set Λ_η in Assumption C can be replaced by Λ_0 . Condition (3.10) holds, at least, in the following cases:

- (i) The linear spaces generated by the cone K^+ are finite-dimensional.
- (ii) The point x_0 is regular with respect to the cone $K_0 = K(\lambda_0)$, $\Lambda = \{\lambda_0\}$.

In case (i), this follows from a compactness argument applied to a sequence $\{\lambda^k\}$ in $\Lambda_{1/k}$. The proof for case (ii) can be found in [16].

Remark 4. Program (Q_t) can be expanded around $t_0 = 0$ as well. That is, we can consider the program

$$(Q'_t) \quad \begin{aligned} & \min_{y \in X} \langle y, Df(x_0) + tD_{xt}^2 f(x_0, 0) \rangle + \frac{1}{2} \langle y, D^2 f(x_0) y \rangle \\ & \text{subject to } g(x_0) + tD_t g(x_0, 0) + D_x g(x_0, 0)y + \\ & \quad \frac{1}{2} t^2 D_{tt}^2 g(x_0, 0) + tD_{xt}^2 g(x_0, 0)y + \frac{1}{2} D_{xx}^2 g(x_0, 0)(y, y) \in K. \end{aligned}$$

Under assumptions similar to those of Theorem 2, an optimal solution $y'(t)$ of (Q'_t) will provide a first-order approximation of $\bar{x}(t) - x_0$.

Existence of the optimal solution $y^*(t)$ (optimal solution $y'(t)$) follows from Assumption C if the program (Q_t) (program (Q'_t)) is convex. In particular, if $D_{xx}^2 g(x_0, 0) = 0$ (for example, if $g(x, 0)$ is linear in x), then $D_{xx}^2 L(x_0, \lambda, 0)$ is equal to $D_{xx}^2 f(x_0, 0)$ for every λ and the constraint mappings of the programs (Q_t) and (Q'_t) are linear in y . Therefore, in this case, existence of $y^*(t)$ and $y'(t)$ is guaranteed by Assumption C.

REFERENCES

- [1] W. ALT, *Lipschitzian perturbations of infinite optimization problems*, Mathematical Programming with Data Perturbations II, A. V. Fiacco, ed., New York, pp. 7–21, 1983.
- [2] ———, *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.
- [3] A. AUSLENDER, *Stability in mathematical programming with nondifferentiable data*, SIAM J. Control Optim., 22 (1984), pp. 239–254.
- [4] A. AUSLENDER AND R. COMINETTI, *First- and second-order sensitivity analysis of nonlinear programs under directional constraint qualification conditions*, Optimization, 21 (1990), pp. 351–363.
- [5] J. F. BONNANS, A. D. IOFFE, AND A. SHAPIRO, Proc. French-German Conf. on Optimization, in Lecture notes in Economics and Math. Systems, D. Pallaschke, ed., Springer-Verlag, to appear.
- [6] J. F. BONNANS, *Directional derivatives of optimal solutions in smooth nonlinear programming*, J. Optim. Theory Appl., 73 (1992), pp. 27–45.
- [7] I. EKELAND AND R. TEMAM, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.
- [8] J. GAUVIN AND R. JANIN, *Directional behavior of optimal solutions in nonlinear mathematical programming*, Math. of Oper. Res., 13 (1988), pp. 629–649.
- [9] F. LEMPPIO AND H. MAURER, *Differential stability in infinite-dimensional nonlinear programming*, Appl. Math. Optim., 6 (1980), pp. 139–152.
- [10] K. MALANOWSKI, *Second order conditions and constraint qualifications in stability and sensitivity analysis of solutions to optimization problems in Hilbert spaces*, Appl. Math. Optim., 25 (1992), pp. 51–79.
- [11] H. MAURER AND J. ZOWE, *First and second-order necessary and sufficient optimality conditions for infinite-dimensional programming problems*, Math. Prog., 16 (1979), pp. 98–110.

- [12] S. M. ROBINSON, *Normed convex processes*, Trans. Amer. Math. Soc., 174 (1972), pp. 127–140.
- [13] S. M. ROBINSON, *Stability theorems for systems of inequalities, Part II: differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.
- [14] ———, *First-order conditions for general nonlinear optimization*, SIAM J. Appl. Math., 30 (1976), pp. 597–607.
- [15] A. SHAPIRO, *Sensitivity analysis of nonlinear programs and differentiability properties of metric projections*, SIAM J. Control Optim., 26 (1988), pp. 628–645.
- [16] ———, *Perturbation analysis of optimization problems in Banach spaces*, Numer. Funct. Anal. Optim., 13 (1992), pp. 97–116.
- [17] C. URSESCU, *Multifunctions with convex closed graph*, Czechoslovak Math. J., 25 (1975), pp. 438–441.
- [18] J. ZOWE AND S. KURCYUSZ, *Regularity and stability for the mathematical programming problem in Banach spaces*, Appl. Math. Optim., 5 (1979), pp. 49–62.

TRACKING AND RESTRICTABILITY IN DISCRETE EVENT DYNAMIC SYSTEMS*

CÜNEYT M. ÖZVEREN† AND ALAN S. WILLSKY‡

Abstract. This paper formulates and analyzes notions of tracking and restrictability for discrete event dynamic systems (DEDS). The DEDS model used is a finite-state automaton in which there is control over some events. A second set of events, called the set of *tracking events*, is also specified, and the tracking problem is one of constructing a compensator so that the tracking event trajectory of the closed loop system follows a given string exactly. This problem is analyzed in detail and, in particular, a characterization of all trackable strings is characterized. The related notion of restrictability is analyzed in which the closed-loop system is required to generate tracking event strings in a given desired language. A relaxed version of this concept is also analyzed, allowing an initial transient before desired language tracking is achieved. Finally, a notion of reliability is introduced and analyzed, which allows for testing if the system can recover from errors in a finite number of transitions, and algorithms are presented for constructing compensators for reliable restrictability. A manufacturing system example is used to motivate and illustrate the problems considered and results obtained.

Key words. tracking, control, reliability, stability, discrete events

AMS(MOS) subject classification. 93

1. Introduction. In the past few years, there has been considerable research on the topic of discrete event dynamic systems (DEDS) [1]–[3], [6]–[9], [18]–[21]. One characteristic of much of this activity is that the control objectives have frequently been stated in linguistic terms, i.e., in terms of characteristics of the possible closed-loop event trajectories. In contrast, in much of our previous work [13], [14], [16], we have focused directly on control concepts of stability, observability, stabilization, and output feedback, providing some of the elements required to develop a regulator theory for DEDS. In particular, to develop such a theory, we need some notion of stability, and the one pursued in [13], [14], [16], which can be considered an error recovery concept, appears to be a natural one in the discrete-event context.

In this paper, we develop another element needed for a regulator theory and which also is much closer to the linguistic concepts explored by others. In particular, we are concerned here with characterizing the *tracking* capabilities of a DEDS in terms of the concept of *trackable languages*, as well as a second notion, *restrictability*, which is a slight generalization of the notion of (language) controllability of Ramadge and Wonham in [19]. While our analysis of restrictability represents a relatively modest addition to the existing theory of controllable languages, we also consider two related, new notions which, we believe, are of some importance, and which are motivated by the desire to introduce notions of stability and error recovery in the theory of DEDS. The first of these concepts is that of eventual or stable restrictability, i.e., the ability to restrict event behavior after a finite start-up period. This would appear to be a useful notion for capturing start-up or mode-switching behavior in DEDS. The second and more involved notion is that of *reliable restrictability*, i.e., the ability of the system to

* Received by the editors September 5, 1989; accepted for publication (in revised form) June 24, 1991. This research was supported by Air Force Office of Scientific Research grant AFOSR-88-0032 and by Army Research Office grant DAAL03-86-K0171. Part of this work was also done while the first author was employed by the Digital Equipment Corporation.

† Telecommunications and Networking, Digital Equipment Corporation, 550 King Street LKG1-2/A19, Littleton, Massachusetts 01460.

‡ Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

return to the desired, restricted behavior following a burst of errors or failures. As we will see, stable restrictability plays a key role in characterizing reliable restrictability.

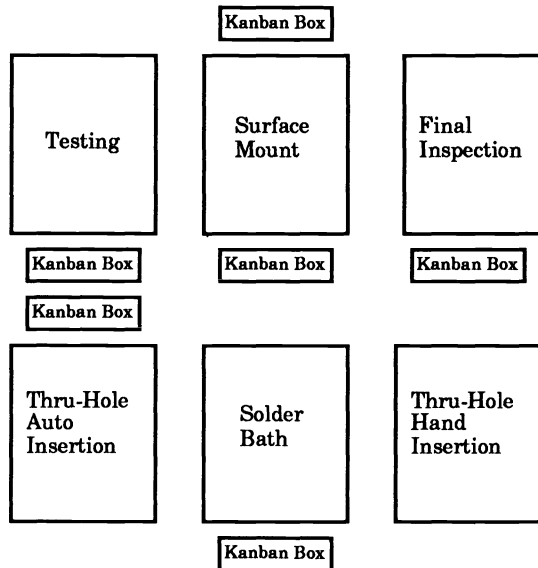


FIG. 1.1. An example of a computer board manufacturing floor.

To motivate the problems considered in this paper and to provide an example that we can use to illustrate their solution, let us briefly describe a particular manufacturing application. More detailed investigations of this and other applications of our regulator theory are given in [15]. Figure 1 illustrates the floor plan of a computer board manufacturing facility consisting of several workstations and capable of soldering surface mount chips on both sides of a board, mixed with thru-hole mounting (via auto-insertion and hand insertion). Another workstation is used for soldering both kinds of thru-hole devices. One workstation is used for testing random board samples at various phases of the manufacturing process, and finally, each board goes through a routine test and inspection after completion. This manufacturing floor uses a Japanese inventory system, termed the Kanban system. Boards are transported through specially marked “kanban” boxes in quantities of 1 to 10 in each box. There are very few kanban boxes between different workstations, guaranteeing that inventories are very low, and thus, among other things, that latency through the manufacturing process is also very low.

A typical board with both sides populated with surface mount components and with mixed thru-hole components goes through the following process: The board first visits the surface mount station for side 1 components, where, first, a solder paste is applied to the board; next, the components are placed on the board; and, finally, solder is applied. The board then goes to the auto-insertion workstation where thru-hole devices are automatically inserted. Next, if necessary, some components are inserted by hand, and the board arrives at the solder bath. There, the boards are first baked, to remove the moisture, and then passed through the wave solder. After that, if there are any side 2 surface mount components, the board goes to the surface mount workstation again for the mounting of side 2 components and, finally, the board goes through final inspection and testing. To construct a manageable example

in the scope of this paper, we capture the dynamics of a system of this type with two workstations and two kinds of boards. We consider a surface mount workstation W1 and a thru-hole workstation W2. One kind of board has only side 1 surface mount components and some thru-hole components. The second kind of board has surface mount components on both sides, as well as thru-hole components. The surface mount workstation will perform two tasks: side 1 mounting (for either kind of board), which we call Task 1, and side 2 mounting, including the inspection for the second kind of board, which we call Task 2. The thru-hole workstation also performs two tasks: thru-hole mounting for the second board, denoted by Task 3, and thru-hole mounting and inspection for the second board, Task 4. Thus, to be completed, the first board must go through Tasks 1 and 4, and the second must go through Tasks 1, 3, and 2, in that order.

Let S_i (respectively, F_i) denote the starting (respectively, finishing) of Task i . In addition, we assume that there is a unit capacity buffer B1 to store boxes going from the surface mount to the thru-hole workstation and another unit capacity buffer B2 in the opposite direction; i.e., there is space for one kanban box in each direction. We then model this system using the automata in Fig. 1.2. Here, circles represent states, and the arcs represent transitions labeled with events. Also, :u indicates that the corresponding event is controllable (i.e., we can decide whether to start a task), and :! indicates that the corresponding event is a “tracking” event, which is identified as an event that is of interest in characterizing desired behavior (see §2 for the precise mathematical model). Suppose that, at a given time, the objective of the manager of such a plant is to manufacture equal amounts of each kind of board. Then, we must perform Task 1 twice, and all the other tasks once to produce one board of each kind. Furthermore, the correct production of these parts requires the correct sequencing of these tasks and the corresponding transfers of boards. In particular, suppose that the time needed to complete Task 1 is comparable to the time to complete Task 4, while the time for Task 3 is comparable the time for Tasks 1 and 2 combined. In this case, we can form a production schedule by performing first Tasks 2 and 1 on the surface mount workstation while performing Task 3 on the other, and then Task 1 on the surface mount workstation while performing Task 4 on the other. Note that it makes no difference if we reverse the order, i.e., require that Tasks 1 and 4 are done first, and so on. In essence, all we must know to construct a schedule is the list of tasks that must be performed and the time it takes to complete each relative to the others. Thus one “cycle” of the schedule, producing one board of each type, corresponds to the completion of any of the sequences in

$$(1.1) \quad L_S = (F_1(F_2F_3 + F_3F_2) + F_2(F_1F_3 + F_3F_1) + F_3(F_1F_2 + F_2F_1))(F_1F_4 + F_4F_1),$$

where multiplication in (1.1) corresponds to concatenation and addition to union (so that $F_1F_2F_3F_4F_1$ and $F_3F_1F_2F_1F_4$ are elements of L_S). Note that by constructing the schedule in this fashion, we are allowing for concurrency; that is, at some points in time, both machines may be working. The control problem then is to exercise the available event controls to ensure that the manufacturing system adheres to the schedule of a succession of sequences from L_S , perhaps with an initial start-up transient and hopefully with the ability to recover gracefully from errors or failures. In this paper, we provide a mathematical framework that allows us to solve problems such as this, and indeed we will revisit this example in later sections to illustrate the construction of controllers that meet design objectives such as this.

In the next section, we introduce our mathematical framework and collect several definitions and results. In §3 we formulate a notion of tracking and present algorithms

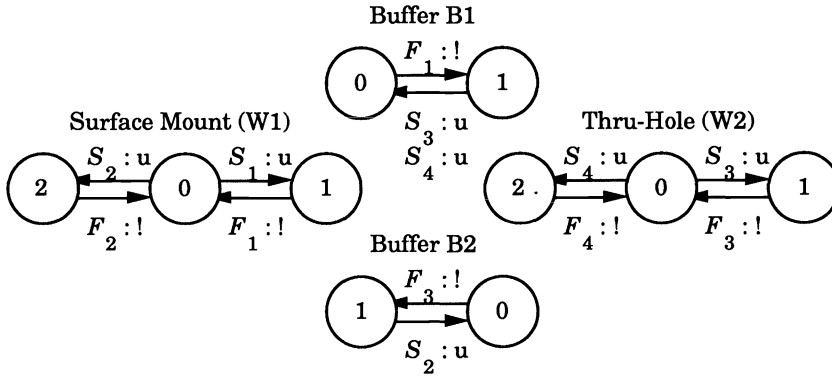


FIG. 1.2. A model of the computer board manufacturing example.

for constructing compensators for tracking specific strings (e.g., particular elements of L_S). In §4 we consider the problem of restricting behavior to a specified set of desired event sequences (e.g., restricting the manufacturing system to successive completion of elements of L_S), a concept very closely related to the notion of controllability of Wonham and Ramadge. Furthermore, in this section, we introduce concepts of eventual and stable restrictability that allow us to address questions concerning the transient behavior of controlled DEDS and investigate the reliability or error-correcting capability of such a system. As we will see, the stronger notion of stable restrictability leads, in general, to considerable computational efficiencies compared to the weaker concept of eventual restrictability. Finally, in §5 we summarize our results and discuss several directions for further work.

2. Background. The class of systems we consider are nondeterministic finite-state automata defined on $G = (X, \Sigma, \Xi, U)$, where X is the finite set of states, with $n = |X|$; Σ is the finite set of possible events; $\Xi \subset \Sigma$ is the set of events that we wish to track; and $U \subset 2^\Sigma$ is the set of admissible control inputs, corresponding to the choices of sets of controllable events that can be enabled. The dynamics defined on G are

$$(2.1) \quad x[k+1] \in f(x[k], \sigma[k+1]),$$

$$(2.2) \quad \sigma[k+1] \in (d(x[k]) \cap u[k]) \cup e(x[k]).$$

Here $x[k] \in X$ is the state, $\sigma[k] \in \Sigma$ is the next event, and $u[k] \in U$ is the next control input. The function $d : X \rightarrow 2^\Sigma$ specifies the set of possible events defined at each state, $e : X \rightarrow 2^\Sigma$ specifies the set of events that *cannot* be disabled at each state, and the function $f : X \times \Sigma \rightarrow 2^X$ is also set-valued. Without loss of generality, we assume that $e(x) \subset d(x)$. Note that in this general framework, there is no loss of generality in taking $U = 2^\Sigma$. Also, by appropriate choice of $e(x)$, we can model situations in which we have enabling/disabling control over some events only at certain states. In parts of the next section, we will use this general framework. In the remainder of this paper, however, we assume the slightly more restrictive framework of [19] in which

there is an event subset $\Phi \subset \Sigma$ such that we have complete control over events in Φ and no control over events in $\bar{\Phi}$, the complement of Φ . In this case, we can take $U = 2^\Phi$ and $e(x) = d(x) \cap \bar{\Phi}$.

The set Ξ , which we term the tracking alphabet, represents events of interest for tracking purposes. This formulation allows us to define tracking over a selected alphabet so that we do not worry about listing intermediate events that are not of direct interest. We use $t : \Sigma^* \rightarrow \Xi^*$, to denote the projection of strings over Σ into Ξ^* . The quintuple $A = (G, f, d, e, t)$ ¹ representing our system can also be visualized graphically as in Fig. 1.2, where the first symbol in each arc label denotes the event. We mark the controllable events by ! and tracking events by !.

We use several basic notions. First, given $Q \subset X$, we use $R(A, Q)$ to denote all the states that can be reached from Q in zero or more steps (so that $Q \subset R(A, x)$). Second, there is the notion of liveness: A DEDS is alive if $d(x)$ is nonempty for all x . We will assume that this is the case. A third notion that we need is the composition $A_{12} = A_1 \parallel A_2$ of two automata $A_i = (G_i, f_i, d_i, e_i, t_i)$, which share some common events. The dynamics of the composition are specified by allowing each automaton to operate as it would in isolation except that when a shared event occurs, it *must* occur in both systems. Note that our manufacturing system can be described by the composition of the four automata in Fig. 1.2, with shared events capturing the fact that a task cannot begin if a board is not available.

Central to our work is the notion of stability studied in [16] (see also [17]). Let E be a given subset of X . We say that a state $x \in X$ is *E-prestable* if every trajectory starting from x passes through E in a bounded number of transitions. The state $x \in X$ is *E-stable* if every state reachable from x is *E-prestable*, and the DEDS is *E-stable* if every $x \in X$ is *E-stable*. Note that *E-stability* for all of A is identical to *E-prestability* for all of A , and that this condition guarantees that all trajectories go through E infinitely often. We refer the reader to [16] for a complete discussion of stability and for an $O(n^2)$ test for *E-stability* of a DEDS.

In [16] we also study state feedback laws of the form $K : X \rightarrow U$, where the resulting closed-loop system is $A_K = (G, f, d_K, e, t)$ with $d_K(x) = (d(x) \cap K(x)) \cup e(x)$. Generally, we wish to avoid feedback laws so that $d_K(x)$ is empty for some x , and we build this constraint into our notions of stabilization. For example, a DEDS is *E-stabilizable* if there exists a feedback K so that A_K is both alive and *E-stable*.

For many control problems, such as those considered in this paper, we must consider compensators that use both current state and event trajectory information. Such a compensator, which is described by a map $C : X \times \Sigma^* \rightarrow U$, yields a closed-loop system A_C , which is the same as A but with

$$(2.3) \quad \sigma[k+1] \in d_C(x[k], s[k]) \triangleq (d(x[k]) \cap C(x[k], s[k])) \cup e(x),$$

where $s[k] = \sigma[0] \cdots \sigma[k]$ with $\sigma[0] = \epsilon$. Note that this class of compensators is similar to the class of supervisors introduced in [19], although, by allowing dependence on the current state, we can achieve a somewhat richer class of behaviors. Note also that we can always write A_C as a DEDS with an expanded (and possibly infinite) state space to realize the dynamics inherent to the map C . As we will see, for our purposes, we can restrict attention to finite state compensators.

In the following, we also use well-known notions of dynamic invariance [16], [18]: A subset Q of X is *f-invariant* if $f(Q, d) \subset Q$, where $f(Q, d) = \bigcup_{x \in Q} f(x, d(x))$. If

¹ On occasion, we will construct auxiliary automata for which we will not be concerned with either control or tracking. In such cases, we will omit e and t from the specification.

$V \subset X$ is f -invariant in A , we denote the restriction of A to V by $A|_V$. We say that a subset Q of X is (f, u) -invariant if there exists a state feedback K such that Q is f -invariant in A_K . However, recall that, in general, we must also preserve liveness. Thus we say that a subset Q of X is a *sustainably (f, u) -invariant* set if there exists a state feedback K such that Q is alive and f -invariant in A_K . Also, given any set $W \subset X$, there is a minimal (f, u) -invariant subset V of W with a corresponding unique *minimally restrictive* feedback K .

Note that, if there exists a cycle in A that consists solely of events that are *not* in Ξ , then the system may stay in this cycle indefinitely. It is not difficult to check for the absence of such cycles, and we assume that this is the case. On occasion, we use the image automaton that keeps track of the state only after the occurrence of tracking events. The state space Y^t of this automaton consists of the union of the set Y_1^t of states that can be reached by tracking events and the set Y_0^t of states to which no events are defined (Y_0^t captures possible start-up behavior). Let $r = |Y^t|$.

It is useful to phrase questions concerning event trajectories in terms of languages [4]. Let L be a regular language over a finite alphabet and let (A_L, x_0) be a minimal recognizer for L . Given $s = pqr$ for some strings p, q , and r over Σ , where p is a prefix of s and r is a suffix of s , we use s/pq to denote r , and we say that q is a substring of s . Finally, we will use the notion of a *complete* language: L is complete if (a) every $s \in L$ is a proper prefix of some other $r \in L$ (so that all trajectories have unlimited extensions), and (b) L is prefix-closed (so that all initial segments of a trajectory are in L). Note that, for a complete language, all strings generated by the recognizer (A_L, x_0) are in L (so that all states are “final” [4]).

3. Tracking. In this section, we first present our notion of tracking and present an algorithm for computing the supremal collection of strings that can be tracked. Later in this section, we present our notion of eventual tracking, which is an extension of our notion of stability. Specifically, we consider the tracking of desired strings after a transient period of a finite number of transitions. For the system A , we will assume the more restrictive framework of Wonham and Ramadge, i.e., that an event controllable at some state is controllable at all the other states. However, various automata that we define in computing trackable strings will belong to the more general framework in [16]. Furthermore, to simplify our presentation of these notions, we will assume that $\Phi \subset \Xi$, i.e., that all controllable events are also in the tracking alphabet.

3.1. Trackable languages. We define tracking as being able to restrict the system behavior so that the automaton starting from the current state must generate a desired string:

DEFINITION 3.1. Given $x \in X$, a string $s \in \Xi^*$ is *trackable from x* if we can find a compensator $C : X \times \Sigma^* \rightarrow U$ such that $t(L(A_C, x)) \subset s\Xi^*$.

As an example, consider the system in Fig. 3.1. Any string in $(\alpha\beta)^*$ is trackable from 0, and a compensator for tracking all such strings can be defined by $C(0, \emptyset) = \{\alpha\}$, $C(1, \alpha) = \{\beta\}$, $C(3, \alpha\gamma) = \{\beta\}$, $C(0, \alpha\beta) = \{\alpha\}$, and so on. As seen in this example, if a string is trackable, then a compensator for tracking it can be constructed easily. Specifically, this compensator should only enable the next event in the string that we wish to track. In the context of manufacturing systems, this notion would be useful in checking if a part can be manufactured at all by the system. In our example, it is obvious that, for example, the second board can be manufactured since the task sequence 1,3,2 can be “tracked.” However, realistically, in a complex system it may not be so obvious if a certain board can be manufactured at all.

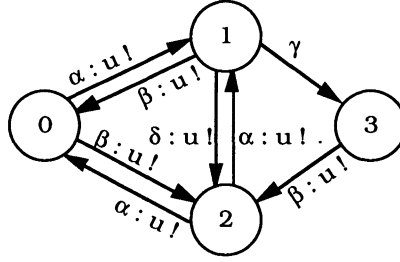


FIG. 3.1. Simple example.

DEFINITION 3.2. A language L is a *trackable language from x* if it is complete and if each string in L is trackable from x .

The class of trackable languages is closed under arbitrary unions, and we let $L_T(A, x)$ denote the supremal language trackable from x . On the other hand, the class of trackable languages is *not* necessarily closed under intersections since the intersection of two complete languages L_1 and L_2 is not necessarily complete, even though it is prefix closed. However, we can construct the supremal complete sublanguage of the intersection. Let the function $\chi : 2^{\Xi^*} \rightarrow 2^{\Xi^*}$ denote removing all the strings, in a given language L , that have no infinite extensions in L , i.e.,

$$(3.1) \quad \chi(L) = \{s \in L \mid s \text{ has an infinite extension in } L\}.$$

Then $\chi(L_1 \cap L_2)$ is a trackable language.

Given some $x \in X$, let us examine the properties of $L_T(A, x)$: First, the first event of a string $s \in L_T(A, x)$ must be defined at some state that is reachable from x by events in $\bar{\Xi}$; i.e., it must be in the set

$$(3.2) \quad d^t(x) = d(R(A|\bar{\Xi}, x)) \cap \Xi.$$

In Fig. 3.1, the first event of a string in $L_T(A, 1)$ may be either β or δ .

Second, the first event of s , say τ , must be trackable from x . We now characterize the set $l_T(x)$ of such events (i.e., the strings of length 1 in $L_T(A, x)$). Let $e^t(x)$ be the set of events in $d^t(x)$ that are either uncontrollable, or events such that, if an event in this set is disabled, then some state in $R(A|\bar{\Xi}, x)$ is no longer alive; see below:

$$(3.3) \quad e^t(x) = (d^t(x) \cap \bar{\Phi}) \cup \{\tau \in d^t(x) \mid \exists y \in R(A|\bar{\Xi}, x) \text{ such that } d(y) = \{\tau\}\}.$$

For example, in Fig. 3.1, $e^t(1) = \{\beta\}$, and e^t is the empty set for all other states. Note that, if $e^t(x)$ contains more than one event, then we cannot track *any* event from x , and if it contains one event, then we can *only* track that event from x . Finally, if $e^t(x)$ is empty, then we can track all events in $d^t(x)$ from x . Thus we have the following proposition.

PROPOSITION 3.3. It holds that $l_T(x) = \{\tau \in d^t(x) \mid \{\tau\} \cup e^t(x) = \{\tau\}\}$.

For example, in Fig. 3.1, $l_T(1) = \{\beta\}$.

After some $\tau \in l_T(x)$ is tracked, the automaton is in some state in $f^t(x, \tau) \triangleq f(R(A|\bar{\Xi}, x), \tau)$. Consequently, the remaining part of the string that can be tracked,

with τ as prefix, must be trackable from *all* these states. Thus we have the following implicit characterization of $L_T(A, x)$.

PROPOSITION 3.4. *It holds that $L_T(A, x) = \bigcup_{\tau \in l_T(x)} \tau \chi(\bigcap_{y \in f^t(x, \tau)} L_T(A, y))$.*

To solve this equation, we construct an automaton $A^t = (G^t, f^t, d^t, e^t, 1)$, where $G^t = (Y^t, \Xi, \Xi, U)$ and 1 is the identity map. For the system in Fig. 3.1, A^t is illustrated in Fig. 3.2. Recall that, if $e^t(x)$ contains one element, then we can only track that event from x . In this case, let

$$(3.4) \quad K'(x) = \begin{cases} \emptyset & \text{if } e^t(x) \neq \emptyset, \\ d^t(x) & \text{otherwise.} \end{cases}$$

In Fig. 3.2, $K'(1) = \emptyset$, since $e^t(1) = \beta$ (thus δ is disabled at state 1), and $K'(0) = \{\alpha, \beta\}$, $K'(2) = \{\alpha\}$. Also, recall that, if $e^t(x)$ contains more than one event, then we cannot track *any* event from x . Let D_T represent such states, i.e.,

$$(3.5) \quad D_T = \{x \in Y^t \mid e^t(x) \geq 2\}.$$

To be able to track complete languages, we must avoid D_T , while preserving liveness. Thus let V be the maximal sustainably (f, u) -invariant subset of \overline{D}_T in A_K^t , and let K_V be the associated minimally restrictive feedback. In Fig. 3.2, V is all of the states. Finally, let $K(x) = K_V(x) \cap K'(x)$. Note that, for $x \in V$, all the events in $d_K^t(x)$ are trackable from x , i.e., $l_T(x) = d_K^t(x)$.

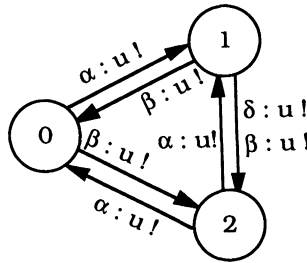


FIG. 3.2. The automaton A^t for Fig. 3.1.

LEMMA 3.5. *If $x \in Y^t \cap \overline{V}$, then $L_T(A, x) = \emptyset$. If $x \in V$, then $L_T(A, x) \subset L(A_K^t, x)$.*

Proof. The proof is straightforward. \square

To compute $L_T(A, x)$, let us first focus on the case in which $A_K^t|V$ is deterministic. In this case, since $l_T(x) = d_K^t(x)$ for all $x \in V$, then, for any $x \in V$, the language generated from x in A_K^t is certainly trackable from x , and, in fact, it is the supremal such language.

PROPOSITION 3.6. *If $A_K^t|V$ is deterministic, then for all $x \in V$, $L_T(A, x) = L(A_K^t, x)$. Furthermore, for all $x \in Y^t \cap \overline{V}$, $L_T(A, x) = \emptyset$.*

Proof. The proof is straightforward using Lemma 3.5 and the fact that, for all $x \in V$ and $\tau \in d_K^t(x)$, $f_K^t(x, \tau)$ is single-valued. \square

To complete the picture when $A_K^t|V$ is deterministic, we must construct $L_T(A, x)$ for the states x in \overline{Y}^t . Let us first seek any such x that is also in $R(A|\Xi, V)$. That

is, there exists $y \in V$ such that x can be reached from y without the occurrence of tracking events. Consider then any $\tau \in l_T(x)$. By definition, $f^t(x, \tau) \subset Y^t$. Furthermore, $f^t(x, \tau) \subset f^t(y, \tau)$. Thus there are two possibilities. Either $f^t(y, \tau)$ is contained in V , or it is not. If it is, then, since K is a minimally restrictive feedback and since $A_K^t|V$ is deterministic, $f^t(x, \tau) = f^t(y, \tau) =$ a single element of V . If $f^t(y, \tau)$ is not contained in V , then the feedback K must disable this event to achieve invariance for V . Thus we must also disable this event at x .² Thus, if we define

$$(3.6) \quad l_V(x) = \{\tau \in l_T(x) | f^t(x, \tau) \in V\},$$

then

$$(3.7) \quad L_T(A, x) = \bigcup_{\tau \in l_V(x)} \tau L_T(A, f^t(x, \tau)),$$

which allows us to compute $L_T(A, x)$ from $L_T(A, y)$, $y \in V$. Next, suppose that $x \notin R(A|\Xi, V)$ and take any $\tau \in l_T(x)$. Again, there are two possibilities: either $f^t(x, \tau) \subset V$ or $f^t(x, \tau) \not\subset V$. Consider the second possibility in which we know that $\tau \notin L_T(A, x)$. There are two cases here: either $\tau \in e^t(x)$ or $\tau \notin e^t(x)$. In the first of these, we cannot disable τ , and thus $L_T(A, x) = \emptyset$. In the latter, we simply disable τ . Consider next the possibility $f^t(x, \tau) \subset V$. There are two cases here as well: either $|f^t(x, \tau)| = 1$ or $|f^t(x, \tau)| > 1$. In the former case, we know that

$$(3.8) \quad L_T(A, x) \supset \tau L_T(A, f^t(x, \tau)),$$

and indeed, if *only* this case occurs, $L_T(A, x)$ is given as in (3.7). However, if $|f^t(x, \tau)| > 1$ for some τ , we have a situation exactly as in the nondeterministic case: essentially, we must intersect the languages $\tau L_T(A, y)$ for all $y \in f^t(x, \tau)$. As this procedure is embedded in the fully nondeterministic case, we describe this case next.

If $A_K^t|V$ is nondeterministic, we first construct a deterministic automaton O^t over subsets of V such that, for each state \hat{x} of O^t , the events defined at \hat{x} are given by the *intersection* of the events defined at each element $x \in \hat{x}$. In particular, we construct an automaton $O^t = (F^t, w^t, v^t)$ over the states 2^V with

$$(3.9) \quad w^t(\hat{x}, \tau) = \bigcup_{x \in \hat{x}} f^t(x, \tau),$$

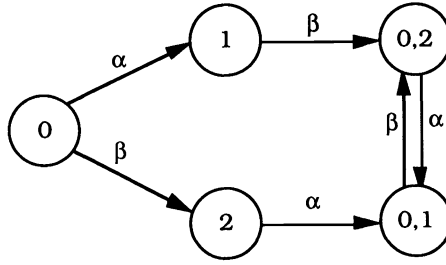
$$(3.10) \quad v^t(\hat{x}) = \bigcap_{x \in \hat{x}} (d^t(x) \cap K(x)) \cup e^t(x).$$

These dynamics can be defined with all of 2^V as the state space. Since we will only use particular initial states, we can restrict attention to the reach of these states under these dynamics. Specifically, we take the state space Z^t of O^t to be

$$(3.11) \quad \begin{aligned} Z^t = & R(O^t, \{\hat{x} \in 2^V | \hat{x} = \{x\} \text{ and } x \in V, \\ & \text{or } \hat{x} = f^t(x, \tau) \text{ for some } x \in \bar{Y}^t, \tau \in l_T(x)\}). \end{aligned}$$

Figure 3.3 illustrates O^t for the automaton in Fig. 3.2. Note that $l_T(3) = \{\beta\}$ and $f^t(3, \beta) = \{2\}$.

² The existence of the feedback K , in fact, guarantees that τ can be disabled while preserving liveness.

FIG. 3.3. The automaton O^t for Fig. 3.2.

Let D_z be the set of dead states in Z^t , i.e.,

$$(3.12) \quad D_z = \{\hat{x} \in Z^t \mid v^t(\hat{x}) = \emptyset\};$$

let Z_V^t be the maximal sustainably (f, u) -invariant subset of $\overline{D_z}$; and let K^t be the associated minimally restrictive feedback. Then, we have the following result, where $Z_s = \{x \mid \{x\} \in Z_V^t\}$.

PROPOSITION 3.7. *Given $x \in Z_s$,*

$$L_T(A, x) = L(O_{K^t}^t, \{x\}).$$

Given $x \in Y^t \cap \overline{Z_s}$,

$$L_T(A, x) = \emptyset.$$

Finally, given $x \in \overline{Y^t}$, let

$$l'_T(x) = \{\tau \in l_T(x) \mid f^t(x, \tau) \in Z_V^t\};$$

then

$$L_T(A, x) = \bigcup_{\tau \in l'_T(x)} \tau L(O_{K^t}^t, f^t(x, \tau)).$$

Note that, if $l'_T(x) = \emptyset$, then $L_T(A, x) = \emptyset$.

The proof of this result is straightforward. Because of the nondeterminism, we must ensure that, for any prefix of a trackable string, the corresponding suffix is trackable from *all* states that can be reached by applying the prefix. The dynamics w^t defined via a union (3.9), and the allowable event function v^t , defined via an intersection (3.10), capture this exactly. A dead state $\hat{x} \in D_z$ then corresponds to a set of states such that *no* event is trackable from all elements of \hat{x} , and thus we must avoid these states and confine the dynamics to Z_V^t . For any singleton element of Z_V^t , i.e., any $\{x\} \in Z_V^t$, it is then easy to compute $L_T(A, x)$. For any other singleton that can be reached by a trackable event, i.e., $x \in Y^t \cap \overline{Z_s}$, we know that the trackable language is empty, since we have started outside of Z_V^t . Finally, for $x \in \overline{Y^t}$, the only trackable events are those that drive x completely within Z_V^t , i.e., $l'_T(x)$, and from there we can compute the suffixes of the trackable strings from x using the

dynamics evolving within Z_V^t . Finally, as we have commented earlier, constructing a compensator for tracking any $s \in L_T(A, x)$ is easy: We just enable the next event that we wish to track, given the string that has already been tracked.

The complexity of computing $L_T(A, x)$ for all x is quadratic in $|Z^t|$. However, as with the cardinality of the state space of an observer [13], $|Z^t|$ may be exponential in $|V|$, and thus computing L_T may have exponential complexity in $|V|$. In [13] we provide some bounds on observer state space size and give examples showing that in many cases the actual observer state space size may be considerably smaller than the worst case exponential bound. Similar analysis can be performed in the present context, and indeed the bounding procedure of [13] can be used to compute a bound on the size of the recurrent part of Z^t . Refer to [10] for an example illustrating both our procedure for computing $L_T(A, x)$ and the worst-case bound using an adaptation of the example used for analogous purposes in [13].

3.2. Eventually trackable languages. A straightforward generalization of the notion of tracking is a notion of tracking a given string in a finite number of transitions. For example, in Fig. 3.1, $(\beta\alpha)^*$ is trackable from state 2 in one transition, namely, after the occurrence of α . We term this a notion of eventual trackability. In the following definition, $(\Xi \cup \{\epsilon\})^{n_t}$ denotes the set of all strings, over Ξ , of length at most n_t , where ϵ denotes the “null” string.

DEFINITION 3.8. Given $x \in X$, a string $s \in \Xi^*$ is *eventually trackable from x* if there exists an integer n_t and a compensator $C : X \times \Sigma^* \rightarrow U$ such that A_C is alive and $t(L(A_C, x)) \subset (\Xi \cup \{\epsilon\})^{n_t} s \Xi^*$. A language L is *eventually trackable from x* if it is complete and if each string in L is eventually trackable from x .

Similar to the class of trackable languages, the class of eventually trackable languages is closed under arbitrary unions, and the supremal complete sublanguage of the intersection of two eventually trackable languages is also eventually trackable. Let $L_{ET}(A, x)$ denote the supremal language eventually trackable from x .

As stated in the following, if a state x is E -prestabilizable for some E , then any string trackable from *all* states in E is eventually trackable from x . For the example in Fig. 3.1, 2 is $\{0, 1\}$ -prestabilizable, and $(\beta\alpha)^*$ is trackable from both 0 and 1.

LEMMA 3.9. *Given $x \in X$ and $E \subset X$ such that x is E -prestabilizable,*

$$L_{ET}(A, x) \supset \bigcap_{y \in E} L_T(A, y).$$

Proof. The proof is straightforward. \square

Conversely, suppose that some string s is eventually trackable from some state x , and let E_s be all the states from which s is trackable. Then x *must* be E_s -prestabilizable, since, otherwise, a trajectory from x may cycle arbitrarily through states from which s is *not* trackable.

PROPOSITION 3.10. *Given x , let \mathbf{E} be the set of all sets $E \subset X$ such that x is E -prestabilizable. Then*

$$L_{ET}(A, x) = \bigcup_{E \in \mathbf{E}} \bigcap_{y \in E} L_T(A, y).$$

Furthermore, for all $x \in X$ and $s \in L_{ET}(A, x)$, $n_t \leq r = Y^t$.

Proof. The proof is straightforward. To prove the second statement, note that n_t can be chosen as the maximum number of tracking events on any trajectory from some $x \in X$ to E . Since r is the cardinality of Y^t , it is an upper bound on n_t . \square

We obtain a slightly tighter formula for $L_{ET}(A, x)$ as follows. Let $Y' \subset X$ (with $r' = |Y'|$) be the set of states from which at least one tracking event is defined, i.e.,

$$(3.13) \quad Y' = \{x \in X \mid d(x) \cap \Xi \neq \emptyset\}.$$

COROLLARY 3.11. *Given x , let \mathbf{E}' be the set of all sets $E' \subset Y'$ such that x is E' -pre-stabilizable. Then*

$$L_{ET}(A, x) = \bigcup_{E' \in \mathbf{E}'} \bigcap_{y \in E'} L_T(A, y).$$

Proof. (\supset) The proof is trivial by the above proposition.

(\subset) Let $s \in L_{ET}(A, x)$, and let $E \subset X$ be a set so that x is E -pre-stabilizable and $s \in L_T(A, x')$ for all $x' \in E$. Next, let

$$E' = R(A \mid \bar{\Xi}, E) \cap Y'.$$

Thanks to our assumption that it is not possible for A to generate arbitrarily long sequences of events in $\bar{\Xi}$, E is E' -prestable. Thus x is E' -pre-stabilizable, and $E' \in \mathbf{E}'$. Also, since all events in $\bar{\Xi}$ are uncontrollable, $s \in L_T(A, y)$ for all $y \in E'$. Therefore $s \in \bigcup_{E' \in \mathbf{E}'} \bigcap_{y \in E'} L_T(A, y)$. \square

To compute \mathbf{E}' , we must check, for each subset E' of Y' , if x is E' -pre-stabilizable. Thus computing $L_{ET}(A, x)$ has complexity exponential in r' . However, testing if a string s is eventually trackable (from some state x) may or may not have exponential complexity, depending on the complexity of the state space of O^t , since all we must do is to compute the set of states in Y' from which s is trackable and test if x is pre-stabilizable with respect to this set. For example, $(\beta\alpha)^*$ is trackable from 0 and 1 in Fig. 3.1. Since 2 is $\{0, 1\}$ -pre-stabilizable, $(\beta\alpha)^*$ is eventually trackable from 2. In fact, both $(\alpha\beta)^*$ and $(\beta\alpha)^*$ are eventually trackable from *all* the states.

4. Restrictability. In this section, we first address the problem of restricting the output behavior of a system to a given language, representing a slight generalization of the notion of controllable languages in the Wonham and Ramadge framework as we also consider arbitrary initial states. Next, we present the concept of eventual restrictability and stable restrictability, which allow us the flexibility of restricting the behavior after a finite number of transitions. Finally, we present and analyze a notion of reliability that allows us to model failure or error events and to test if the system can be made to recover following the occurrence of a burst of errors. Throughout this section, we consider the general setting in which Φ need not be contained in Ξ .

4.1. Basic notion. Given a complete language L over Ξ and a state x , our notion of restrictability is defined as the ability to control the system so that all the trajectories generated from x in the closed-loop system are in L .

DEFINITION 4.1. Given $x \in X$ and a complete language L over Ξ , x is *L -restrictable* if there exists a compensator $C : X \times \Sigma^* \rightarrow U$ such that the closed-loop system A_C is alive and $t(L(A_C, x)) \subset L$. Given $Q \subset X$, Q is *L -restrictable* if all $x \in Q$ are L -restrictable. Finally, A is *L -restrictable* if X is L -restrictable.

The class of L -restrictable sets is closed under arbitrary unions and intersections. Let X_L denote the maximal L -restrictable set. To compute X_L , we first construct a recognizer for L and then formulate the problem of restrictability as one of stabilizability of the composite of this recognizer and A . In the rest of this section, we present this approach and establish connections to the work of Wonham and Ramadge.

Let (A_L, x_0) be a minimal recognizer for L and let Z_L denote its state space. Let A'_L be an automaton that is the same as A_L , except that its state space is $Z'_L = Z_L \cup \{b\}$, where b is a state used to signify that the event trajectory is no longer in L . Also, we let $d'_L(x) = \Xi$ for all $x \in Z'_L$, and

$$(4.1) \quad f'_L(x, \sigma) = \begin{cases} f_L(x, \sigma) & \text{if } x \neq b \text{ and } \sigma \in d_L(x), \\ \{b\} & \text{otherwise.} \end{cases}$$

As an example, consider the system illustrated in Fig. 4.1(a), which is identical to an example in [22]. We have two simple automata, each of which can be thought of as a machine in a manufacturing system. Each of these machines has two states so that state 0 corresponds to being idle, and 1 corresponds to working on a part. Event α (respectively, δ) signifies that the first (respectively, second) machine started working, and event β (respectively, γ) signifies that the first (respectively, second) machine is finished with the part. Events α and δ are assumed to be controllable. Their composition, which models all the behavior that can be generated by the two machines, is illustrated in Fig. 4.1(b). Suppose that the first machine feeds the second one (i.e., after the first machine is finished with a part, the second one starts working on it), and suppose that there is a buffer of size one between the two machines. Our goal is to design a compensator such that the buffer never overflows; i.e., at any given time, there can be at most one part in the buffer. This implies that the set of strings that we wish to allow must have β and δ alternate. A recognizer for this language, and, in fact, the automaton A'_L with the initial state 0, is illustrated in Fig. 4.2, where we have taken $\Xi = \{\beta, \delta\}$ as the tracking alphabet.

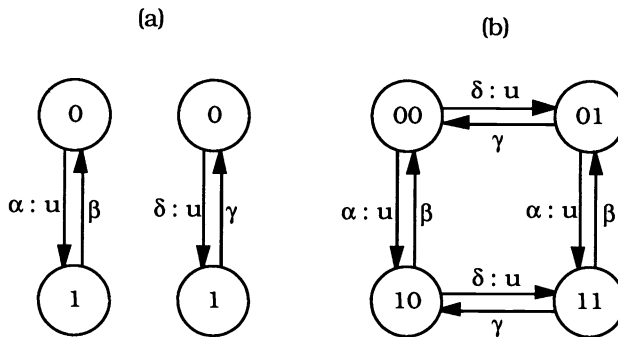
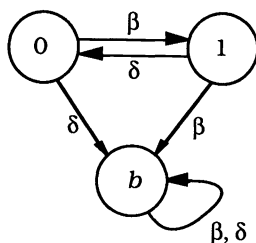
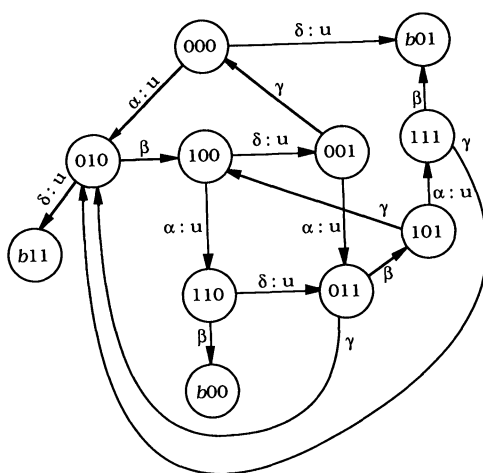


FIG. 4.1. Example for restrictability.

Let A^{LA} denote the composite $A'_L \parallel A$ and let

$$(4.2) \quad E^{LA} = \{(x, y) \in X^{LA} | x \neq b\}.$$

For example, Fig. 4.3 denotes the composite of the automata in Figs. 4.1(b) and 4.2, where the first component of the labels of each state represent the state of A'_L , the last two represent the state of A , and transitions defined at states with the first component equal to b have been ignored for simplicity. Note that E^{LA} is the set of all states that do *not* have b as their first component.

FIG. 4.2. Automaton A'_L .FIG. 4.3. Composite of A and A'_L .

Given $Q \subset X$, let $I(Q)$ denote the maximal sustainably (f, u) -invariant subset of Q and let K_I denote the associated minimally restrictive feedback. Then we have the following proposition.

PROPOSITION 4.2. *A state $x \in X$ is L -restrictable if and only if $(x_0, x) \in I(R(A^{LA}, (x_0, x)) \cap E^{LA})$. Furthermore, a compensator for restricting the behavior of x to L can be constructed using the closed-loop automaton $A_{K_I}^{LA} = (G', f', d')$ and the*

initial state (x_0, x) as follows:

$$C(y, s) = \begin{cases} d'((x_0, x)) & \text{if } s = \epsilon, \\ d'(f'((x_0, x), s)) & \text{if } s \in L(A_{K_I}^{LA}, (x_0, x)), \\ \text{don't care} & \text{otherwise.} \end{cases}$$

Proof. The proof is straightforward by assuming the contrary in each direction. \square

In Fig. 4.3, if we disable δ at 000 and 010, and α at 100 and 101, then we see that all the states of A are L -restrictable.

The compensator C is implemented as follows: Given the initial state, x , of A , we initiate $A_{K_I}^{LA}$ at (x_0, x) . The compensator is simply the set-valued function of the present state of $A_{K_I}^{LA}$ given in Proposition 4.2.

Finally, the following result presents a straightforward construction for X_L .

PROPOSITION 4.3. *We have that $X_L = \{x \in X | (x_0, x) \in I(R(A^{LA}, S_L))\}$, where $S_L = \{(x_0, x) \in X^{LA}\}$, and the complexity of this computation is $O(|X^{LA}|^2)$.*

Proof. The proof is straightforward. Since the complexity of computing $I(Q)$ is quadratic in the cardinality of the state space, the total complexity is $O(|X^{LA}|^2)$ (see [16]). \square

In our board manufacturing example, our objective is to follow the specified schedule that corresponds to restricting system behavior to L_{BM} , the prefix closure³ of L_S^* . It is not difficult to check that some of the states of the system of Fig. 1.2 are restrictable with respect to L_{BM} . In particular, let the quadruple of the states of W1, W2, B1, and B2 represent the state of the composite system. Then Fig. 4.4(a) represents the closed-loop system after restricting the behavior from state $(0, 0, 1, 1)$. (For simplicity in this figure, intermediate states are not shown explicitly, but the end of each transition terminates at a state.) Note, for example, that initially, only S_2 is enabled, since enabling S_1 could lead to F_1 , which would overflow B1, and enabling S_3 could lead to F_3 , which would overflow B2.

We can now relate our results to the notion of controllable languages of Wonham and Ramadge. We refer the reader to [19] for definitions. Specifically, let all events be tracking events (i.e., let $\Xi = \Sigma$), let L be the specified legal language, and let some $x \in X$ be the given initial state of A . Then Proposition 4.4 follows.

PROPOSITION 4.4. *$L(A_{K_I}^{LA}, (x_0, x))$ is the supremal controllable sublanguage of the legal language L .*

Proof. This is straightforward to check from the definitions in [19] and the fact that K_I is minimally restrictive. \square

As an example, if the initial state of the system in Fig. 4.1(b) is 00, then the supremal controllable sublanguage of L is the language generated by state 000 in Fig. 4.3 with δ disabled at 000 and 010, and α disabled at 100 and 101, as before. This compensator is also the same as the one computed in [22] for this example.

As a final comment, note that, from the development in §3, it might be expected that we would have presented results on “maximal” or “minimal” restrictable languages. These concepts, however, are trivial: The maximal language to which we can restrict behavior is obviously Ξ^* , while a number of minimal restrictable languages are possible. For example, if $e(x) \neq \emptyset$, disable all controllable events at this state;

³ This allows for the fact that we may be in the *middle* of one of the sequences in L_S in (1.1), which certainly is consistent with our desire to follow the schedule.

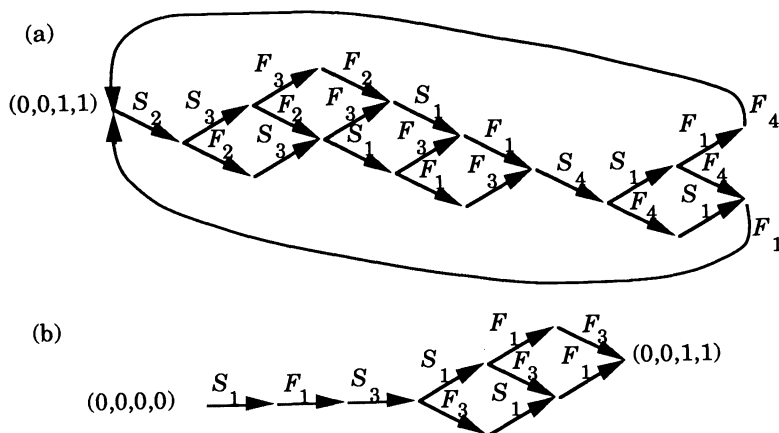


FIG. 4.4. Part of the closed loop system for the compensated board manufacturing example.

if $e(x) = \emptyset$, disable all but one controllable event. Thus, in this context, it is more meaningful to fix L and consider the questions we have addressed here.

4.2. Eventual restrictability and stable restrictability. As noted in the preceding section, some of the states in the manufacturing system of Fig. 1.2 are L_{BM} -restrictable. Others (such as $(0,0,0,0)$) are not. However, for such states, it is possible to design control rules so that we do begin to follow the desired schedule after a short initial set-up transient. This provides the motivation for a natural generalization of our notion of restrictability. For example, consider the system in Fig. 4.5, where $\Xi = \Sigma$, and suppose that $L = (\alpha\gamma + \beta\delta)^*$. The automaton A'_L is illustrated in Fig. 4.6 and the automaton A^{LA} is illustrated in Fig. 4.7, where the transitions defined at state $b0$ have been ignored for simplicity. Note that 0 is L -restrictable, whereas 1 is *not*. However, if the system starts in state 1, the next transition takes state 1 to state 0, and the language generated from that point on can be restricted to L . We term this eventual restrictability.

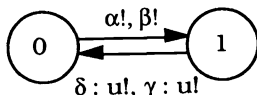
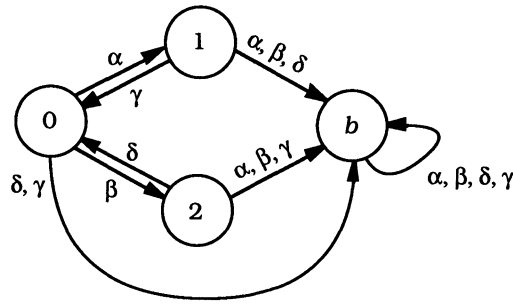
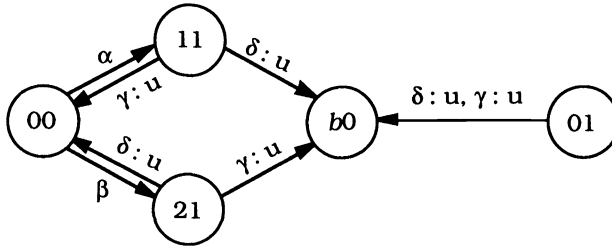


FIG. 4.5. Example for eventual restrictability.

DEFINITION 4.5. Given $x \in X$ and a complete language L over Ξ , x is *eventually L -restrictable* if there exists an integer n_a and a compensator $C : X \times \Sigma^* \rightarrow U$ such that the closed-loop system A_C is alive and $t(L(A_C, x)) \subset (\Xi \cup \{\epsilon\})^{n_a} L$. Given $Q \subset X$, Q is *eventually L -restrictable* if all $x \in Q$ are eventually L -restrictable. Finally, A is *eventually L -restrictable* if X is eventually L -restrictable.

FIG. 4.6. Automaton A'_L for $L = (\alpha\gamma + \beta\delta)^*$.FIG. 4.7. Composite of A and A'_L for the eventual restrictability example.

The class of eventually L -restrictable sets is closed under arbitrary unions and intersections, and thus it has a maximal element X_{EL} . A set closely related to X_{EL} is X_{SL} , the maximal X_L -prestable set, i.e., the set of states that can be driven into states from which L -restrictability can be achieved. The advantage of considering X_{SL} is that it is easy to compute and (directly from the results on prestabilizability [16]) for states in X_{SL} , $n_a \leq r$ and the computation of X_{SL} has complexity $O(n^2)$.

DEFINITION 4.6. Given $x \in X$ and a complete language L over Ξ , x is *stably L -restrictable* if x is X_L -prestable. Given $Q \subset X$, Q is *stably L -restrictable* if all $x \in Q$ are stably L -restrictable. Finally, A is *stably L -restrictable* if X is stably L -restrictable.

A compensator for stable restrictability can be constructed by using two compensators in tandem: The first one is a state feedback that prestabilizes A with respect to X_L . The second one is the compensator of Proposition 4.2 for restricting the language generated by x to L , where x is the element of X_L that the trajectory first visits.

One natural question that arises concerns the relationship between X_{EL} and X_{SL} . Clearly, $X_{EL} \supset X_{SL}$, and, in fact, for many systems and languages the two sets are equal (in particular, this is true if A is stably L -restrictable). For example, it can be verified that our computer board manufacturing system is stably L_{BM} -restrictable, and Fig. 4.4(b) illustrates part of the closed-loop system that ensures that $(0, 0, 0, 0)$

reaches $(0, 0, 1, 1)$ in a finite number of transitions. However, as first shown in [5],⁴ there are systems and languages for which $X_{EL} \neq X_{SL}$, and, in some of these cases, the length n_a of the initial transient until we begin strings in L need not be bounded by r and can be quite long.

While it is beyond the scope of this paper to present a full investigation of the relationship between X_{EL} and X_{SL} and conditions under which they are equal (or at least n_a is small), we can make a few remarks concerning this issue. A simple example adapted from [5] is given in Fig. 4.8, where all events are tracking events and no event is controllable. If we let $L = \delta^* + \alpha^k \alpha^* \beta \delta^*$ for some fixed but arbitrary integer k , then $X_L = \{1\}$, $X_{EL} = \{0, 1\}$, $X_{SL} = \{1\}$, and $n_a = k$. Note that, if L were taken as δ^* , then $X_{EL} = X_{SL} = \{1\}$, while, if L were $\alpha^k \alpha^* \beta \delta^*$, then X_{EL} and X_{SL} are both empty. The difficulty thus appears to be related to the interaction between the two components of L together with the long prefix $\alpha^k \alpha^* \beta$ of one of these components. Some of these difficulties are removed if we restrict our attention to a subclass of languages corresponding to the successive completion of a sequence of “tasks,” i.e., to languages of the form $(L_f)^{*c}$, the prefix closure of the language of all concatenations of strings in the finite set $L_f = \{w_1, \dots, w_m\}$ (note that this is exactly the form of $L_{BM} = (L_S)^{*c}$ for our manufacturing example).

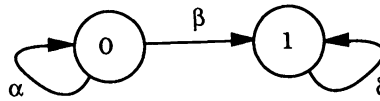


FIG. 4.8. Example illustrating the bound on n_a for $L = \delta^* + \alpha^k \alpha^* \beta \delta^*$.

Restricting ourselves to languages of this form does eliminate the situation depicted in Fig. 4.8, but it is not sufficient to guarantee that $X_{EL} = X_{SL}$. For example, consider the system in Fig. 4.9, where all events are tracking events and no event is controllable. If $L = (L_f)^{*c}$ with $L_f = \{\alpha\beta, \beta\alpha, \alpha\delta, \beta\delta, \mu\}$, then $n_a = 1$ and $X_L = X_{SL} = \{0, 3\}$, but $X_{EL} = \{0, 1, 2, 3\}$. One of the difficulties in this case is that there are ambiguities in the parsing of strings that are eventually in L . For example, the string $\alpha\beta\alpha\beta$ can be given the following two parsings: (1) two occurrences of $w_1 = \alpha\beta$; or (2) an initial prefix of α , an occurrence of $w_2 = \beta\alpha$, and the initiation β of either another occurrence of w_2 or an occurrence of $w_4 = \beta\delta$. While a complete answer to the constraints on L_f under which $X_{EL} = X_{SL}$ for $L = (L_f)^{*c}$ remains open, there are some sufficient conditions that guarantee this. We present here one such condition, which, on the one hand, is more restrictive than necessary, but, on the other hand, is easily interpreted and should not be an unreasonable assumption in many applications.

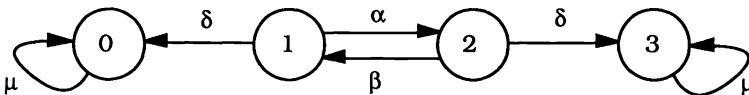


FIG. 4.9. Example illustrating unequivalent X_{EL} and X_{SL} .

⁴ We are grateful to the authors of [5] for pointing out this subtlety to us.

PROPERTY 4.7. *A set of strings L_f has the unique parsing property if, for any string $s \in \Xi^*$, either s possesses no substring that is an element of L_f , or there is a unique way in which to write*

$$s = p_1 w_{i_1} p_2 w_{i_2} \cdots p_m w_{i_m} p_{m+1},$$

where w_{i_1}, \dots, w_{i_m} are (not necessarily distinct) elements of L_f , and none of the strings p_1, \dots, p_{m+1} contains a substring that is an element of L_f .

Note that, in general, this property need not be an easy one to check. There are, however, a number of important necessary conditions for this property to hold. In particular, no element of L_f can be a substring of another, and no prefix of one element of L_f can be a suffix of another (so that no word can be a cyclic permutation of another). Note further that this condition does *not* hold for our manufacturing example, since the string $F_1 F_2 F_3 F_4 F_1 F_2 F_3 F_4 F_1$ can be thought of either as the word $F_1 F_2 F_3 F_4 F_1$ in L_S followed by the string $F_2 F_3 F_4 F_1$ or as the string $F_1 F_2 F_3 F_4$ followed by the word $F_1 F_2 F_3 F_4 F_1$ in L_S . One simple condition under which this property does hold is if there is either a unique special element of Ξ that only appears at the end of each element of L_f , indicating “task completion” (or, equivalently, a special element that appears only at the start of each element of L_f , indicating “task initiation”). For example, in our manufacturing example this would correspond to a simple modification to include explicitly a final event in each element of L_S corresponding to transferring the completed pair of boards to the final inspection station.

PROPOSITION 4.8. *Let $L = (L_f)^*{}^c$, where $L_f = \{w_1, \dots, w_m\}$ has the unique parsing property. Then $X_{EL} = X_{SL}$.*

Proof. As shown in [5], there is in general a (very large) upper bound on n_a and a finite-state compensator C such that $t(L(A_C, x)) \subset (\Xi \cup \{\epsilon\})^{n_a} L$ for all $x \in X_{EL}$. Suppose that $X_{EL} \neq X_{SL}$ and take any $x_0 \in X_{EL} \setminus X_{SL}$, the complement of X_{SL} in X_{EL} . Thanks to unique parsing, any state reachable (in A_C) from a state in X_{EL} must also be in X_{EL} , so that $f(x_0, s) \in X_{EL}$ for any $s \in L(A_C, x_0)$. Furthermore, since $x_0 \notin X_{SL}$, we can find a state path of arbitrarily long length beginning at x_0 that does not enter X_{SL} . By the finiteness of the composition of A and C , there then must exist a string in $L(A_C, x_0)$ that produces a cycle in the composite that stays in $X_{EL} \setminus X_{SL}$. That is, we can find a string p and s so that $ps^* \subset L(A_C, x_0)$, corresponding to a path completely within $X_{EL} \setminus X_{SL}$, where s represents the string of events around the cycle. Since $x_0 \in X_{EL}$ and since n_a is bounded, for k sufficiently large, $t(ps^k)$ contains a suffix in L . Let us first examine the case when $t(s)$ does not contain a substring in L_f . Then, thanks to eventual restrictability and to the finite length of words in L_f , we know that, for k sufficiently large, we must encounter strings of the form

$$(4.3) \quad ps^k = p \cdots u_1 w_{i_1} w_{i_2} \cdots w_{i_m} v_{m+1},$$

where

- (i) $w_{i_j} = v_j s^{n_j} u_{j+1}$,
- (ii) $t(w_{i_j}) \in L_f$,
- (iii) n_j is a nonnegative integer,
- (iv) $s = u_i v_i$ for $i = 1, \dots, m$

(see Fig. 4.10). Since s is a finite string and L_f is a finite set, it must be that, and such that, for some k and some $l < m$,

$$(4.4) \quad w_{i_l} = w_{i_m}.$$

In this case, since $v_m = v_l$ and $s = u_l v_l$, we see that

$$(4.5) \quad w_{i_l} w_{i_{l+1}} \cdots w_{i_{m-1}} = v_l s^{n_l} u_{l+1} \cdots v_{m-1} s^{n_{m-1}} u_l = \sigma^N,$$

where $\sigma = v_l u_l$ and $N = n_l + \cdots + n_{m-1} + m - l$. Note that (1) σ is simply a cyclic permutation of s , so that $\sigma \in L(A_C, y)$ for some $y \in X_{EL} \setminus X_{SL}$ on the cycle, and (2) $t(\sigma^N)$ is precisely a concatenation of strings in L_f . In the other case in which s contains a substring in L_f , we can still obtain a parsing as in (4.3) with the second condition changed to the statement that there exists a finite integer $r > 0$ so that $t(w_{i_j}) \in (L_f)^r$ (here we can take r as any integer greater than the ratio of the length of s to the length of the shortest element of L_f). Then, because of the finiteness of $(L_f)^r$, we can again find $l < m$, so that (4.4) and (4.5) hold, and thus so that the same two conditions hold for σ and y .

$$p \quad \dots \quad \overbrace{u_1 v_1}^s s^{n_1} \overbrace{u_2 v_2}^s s^{n_2} \overbrace{u_3 v_3}^s \dots$$

$\underbrace{\hspace{10em}}_{w_{i_1}} \quad \underbrace{\hspace{10em}}_{w_{i_2}}$

FIG. 4.10. The parsing of ps^k .

Consider then this state y . Since $y \notin X_L$, there is some string $d \in L(A_C, y)$ such that $t(d)$ is not in L . However, $(\sigma^N)^* d \subset L(A_C, y)$. By unique parsing, since σ^N is a sequence of elements of L_f , $t[(\sigma^N)^k d]$ cannot be an element of L for any k , since $t(d) \notin L$. On the other hand, since $y \in X_{EL}$ such strings *must* be in L as they are the suffixes of words of arbitrarily long length in $t[(\sigma^N)^* d]$. This establishes a contradiction to the assumption that $X_{EL} \setminus X_{SL}$ is nonempty. \square

As we have indicated, the complete characterization of the relationship between X_{EL} and X_{SL} remains open. Several other less restrictive conditions on the structure of L are known that guarantee $X_{EL} = X_{SL}$, but none of these appear to have as simple and reasonable interpretation as Property 4.7. Completely open is the question of conditions on the automaton A (rather than the language L) that guarantee $X_{EL} = X_{SL}$. However, while these represent interesting research questions, it is our opinion that X_{SL} is a more meaningful object to begin with, since in essence for $x \in X_{EL} \setminus X_{SL}$ eventual restrictability happens in a sense by “accident.” In contrast to $x \in X_{SL}$, we explicitly drive the system to X_L at which point we *know* that generation of strings in L commences. Thus, as briefly discussed in the §5, it is the notion of stable restrictability that plays a central role in [11] in our development of a theory of hierarchical aggregation based on the concept of task completion.

4.3. Reliability. Our final generalization of restrictability is very similar to the notion of resiliency introduced in [12]. Specifically, we allow a set of failure events and require that following a burst of failures, the system generates strings in L within a finite number of transitions after the burst ends. For example, in our manufacturing system, suppose that parts are detected to be defective as a kanban box arrives at W2. In particular, suppose that this “failure event” happens immediately after S_2 occurs from state $(0,0,1,1)$ in Fig. 4.4(a). Then we would essentially observe a transition to state $(2,0,0,0)$, since B1 suddenly becomes empty while W2 is still idle. To start production again, our goal at this point is to reach state $(0,0,1,1)$. Note that we can

stabilize $(2,0,0,0)$ with respect $(0,0,1,1)$, since all we must do is wait until T_2 finishes and the state transitions to $(0,0,0,0)$, which we know is stable. Thus, in this case, we can recover from the initial failure within a few steps. We capture this recovery procedure as follows: To be consistent with our current framework, let us decompose Ξ into tracking events Ξ_t and failure events Ξ_f (instead of defining a new alphabet). A natural assumption is that no event in Ξ_f is controllable (since, otherwise, we can just disable them). Given an integer $i \geq 1$ and $s \in L(A, x)$ for some $x \in X$, we say that s is a *failure sequence with at most i failures* if both the first and the last events of s are in Ξ_f and *at least one but at most i events of s are in Ξ_f* . We define reliability as follows. (We build this notion on the notion of stable restrictability.)

DEFINITION 4.9. Given $x \in X$, a complete language L over Ξ_t , and an integer $i \geq 1$, x is *i -reliably L -restrictable* if x is stably L -restrictable in $A|\bar{\Xi}_f$ and there exists a compensator $C : X \times \Sigma^* \rightarrow U$ such that the closed-loop system $A_C|\bar{\Xi}_f$ is alive. Also, for all failure sequences $s \in L(A_C, x)$ with at most i failures, $f(x, s)$ is stably L -restrictable in $A_C|\bar{\Xi}_f$. Given $Q \subset X$, Q is *i -reliably L -restrictable* if all $x \in Q$ are i -reliably L -restrictable. A is *i -reliably L -restrictable* if X is i -reliably L -restrictable.

The class of i -reliably L -restrictable sets is closed under unions and intersections. Let X_R^i denote the maximal i -reliably L -restrictable set, and let $X_R^\infty \triangleq \bigcap_{i=1}^\infty X_R^i$. Note that $X_R^0 = X_{SL}$, where X_{SL} is defined for $A|\bar{\Xi}_f$. The following proposition is immediate.

PROPOSITION 4.10. *The sets X_R^i are nested, i.e.,*

$$X_R^{i+1} \subset X_R^i,$$

and, if $X_R^{i+1} = X_R^i$, then $X_R^j = X_R^i$ for all $j \geq i$ including ∞ .

It remains to describe a recursive procedure for computing X_R^i beginning from X_R^0 . Let Y^i , for integers $i \geq 0$, denote the set of states x such that either *no* failure events are defined at x or all the failure events take x to a state in X_R^i , i.e.,

$$(4.6) \quad Y^i = \{x \in X | d_f(x) = \emptyset \text{ or for all } \sigma \in d_f(y), f(y, \sigma) \in X_R^i\},$$

where $d_f(x) = d(x) \cap \Xi_f$. Note that $Y^{i+1} \subset Y^i$.

Consider then what it means for a state $x \in X$ to be 1-reliably L -restrictable. First, we must have that $x \in X_R^0$. Second, we must have that any state that can be reached from x with one failure event must be stably L -restrictable with failure events turned off. To be precise, define

$$(4.7) \quad L_1(A, x) = \{s \in L(A, x) | \text{only the last element of } s \text{ in } \Xi_f\}.$$

Thus $L_1(A, x)$ are the possible event trajectories leading up to and including the first failure when we start in x . Then we must have, for any $s \in L_1(A, x)$, that $f(x, s) \in X_R^0$. Note that this implies that all of the states along any trajectory from x to $f(x, s)$ must lie completely in Y_0 . Thus let X^0 denote the maximal sustainably (f, u) -invariant set in Y^0 and let K^0 denote the associated feedback. Then we have Lemma 4.11.

LEMMA 4.11. *We have that $X_R^1 = R^0$, the maximal stably L -restrictable subset of X^0 in $A_{K^0}|\bar{\Xi}_f$.*

Proof. That X_R^1 is contained in R^0 is clear from the preceding argument. To show the opposite inclusion, take any $x \in R^0$. Then we can find a compensator such that, with x as initial state, only strings in L are allowed in the closed-loop system (with failure events turned off), and the trajectories stay in Y^0 . Then, following a failure

event, the system can only make a transition to a state y that is stably restrictable, and thus we can restrict the language generated from y to L within a finite number of transitions. Therefore x is 1-reliably L -restrictable. \square

Note that, from the argument preceding the lemma statement, we might conclude that X_R^1 is simply $X_R^0 \cap X^0$. However, in making X^0 invariant, we have applied a feedback K^0 , and this may then restrict what further feedback can be applied to achieve stable restrictability. Thus, in general, X_R^1 may be smaller than $X_R^0 \cap X^0$.

Continuing with our construction, let X^i denote the maximal sustainable (f, u) -invariant subset in Y^i of $A|\Xi_f$, and let K^i be the associated state feedback. Note that, because of the nesting of the Y^i , K^i is compatible with K^{i-1} (i.e., any event disabled by K^{i-1} is also disabled by K^i). We then have the following proposition.

PROPOSITION 4.12. X_R^{i+1} is the maximal stably L -restrictable subset of X^i in $A_{K^i}|\Xi_f$.

Proof. The proof is similar to the proof of Lemma 4.11. \square

Thus the full recursive procedure is the following: (1) Compute X_R^0 using the stable restrictability results of the preceding section applied to $A|\Xi_f$; (2) Given X_R^i , compute Y^i from (4.6); (3) Compute X^i and K^i using the (f, u) -invariance results discussed in §2; (4) Compute X_R^{i+1} using the stable restrictability results applied to $A_{K^i}|\Xi_f$. Also, as a byproduct of the above construction, we obtain the following result.

COROLLARY 4.13. It holds that $X_R^\infty = X_R^i$ for some $i \leq |Y'|$, where $Y' = \{x | d_f(x) \neq \emptyset\}$.

Thus we can compute X_R^∞ in a finite number of steps, and, in fact, the complexity of this computation is $O(|Y'| |X^{LA}|^2)$.

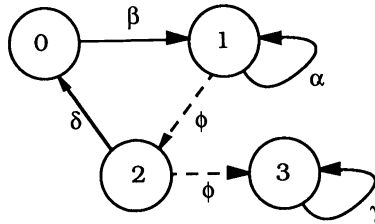


FIG. 4.11. Reliable restrictability example: $\Xi_t = \{\alpha, \beta, \delta, \gamma\}$, $\Phi = \emptyset$, $\Xi_f = \{\phi\}$.

As an example, consider the system in Fig. 4.11, where $\Xi_t = \{\alpha, \beta, \delta, \gamma\}$, $\Phi = \emptyset$, and $\Xi_f = \{\phi\}$. Let $L = \beta\alpha^*$; then $X_R^0 = \{0, 2\}$ and $Y^0 = \{0, 1, 3\}$. Thus $X^0 = Y^0$, and K^0 is a trivial feedback that enables all events. Also, $X_R^1 = \{0\}$. Thus state 0 can recover from a single failure. However, if the failure ϕ occurs at state 2, then a transition is made to state 3 that is not stably L -restrictable. Continuing, we obtain $X_R^2 = \emptyset$. Thus state 0 cannot recover from 2 or more failures.

5. Conclusions. In this paper, we have investigated notions of tracking, restrictability, and reliability for discrete-event dynamic systems. We have developed algorithms for constructing trackable languages, testing restrictability and reliability, and constructing compensators for stable and reliable restriction of system behavior. As we have illustrated, the concepts arise naturally in operation sequence control in

flexible manufacturing systems, and we expect that they will also prove to be relevant in a number of other contexts as well. The work in this paper complements our stability analysis in [16] in that the notions of eventual trackability and eventual restrictability lead to particular choices for the set E that we use for stability. In the case of partial observations, our results in this paper can be combined with our results on stabilization by output feedback in [14] to address problems of tracking and restrictability in the context of intermittent observations of events. As we have shown in [13] and [14], problems of stabilization by output feedback have polynomial complexity if the observer state space is also polynomial. Since our conditions for restrictability are also based on stabilizability and since we have seen how to place the problem of controllable languages of Wonham and Ramadge in our framework, we see that the reason behind the NP-completeness of this problem [20] in the case of partial observations is the cardinality of the observer state space. Thus, if, in fact, the observer has polynomial state space (as it does in many cases [13]), then the problem of controllable languages for the case of partial observations can also be solved in polynomial time.

Another major problem of computational complexity in DEDS arises in the case of interacting automata. If, for example, we have m interconnected subsystems each with n states, then their composition may have n^m states. In this case, it would be extremely worthwhile to develop methods for obtaining aggregate models for each subsystem before addressing higher-level problems involving their interconnection. For example, consider again the manufacturing system of Fig. 1.2. Obviously, the “event” F1, corresponding to side 1 mounting, involves a sequence of commands for the surface mount workstation W1 (indeed, there may be several such sequences corresponding to mounting several different parts). Thus, at a lower level, we see that we have a restrictability problem for the control of each of the machines in Fig. 1.2, and this figure represents a higher-level version of each component system, aggregated to a level appropriate for the consideration of multi-machine coordination. An obvious question, then, concerns the problem of constructing higher-level models as in Fig. 1.2 from lower-level descriptions. In [11] we use the notions of restrictability and stable restrictability presented in this paper to develop such an hierarchical aggregation procedure based on the idea of transforming restricted event sequences at a lower level to single “task” events at higher levels. Obviously, we can also imagine performing such an aggregation procedure at a number of scales. For example, suppose that we have a set of schedules corresponding to different production operating points corresponding to distinct percentage of mixes of several computer boards. We can then construct compensators for implementing each, and eventual or stable restrictability will provide us with the means of changing the set-up from one schedule to another. Thus we can construct a higher-level model based on the set of all schedules by combining the respective compensators for each. Each occurrence of a higher-level event in this model would correspond to completing a cycle of some schedule, i.e., completing a certain number of each type of board. Then the plant manager could try to meet the actual demand distribution by switching between appropriate schedules based on this aggregate, higher-level model capturing operating behavior for all schedules.

REFERENCES

- [1] P. E. CAINES AND S. WANG, *Classical and logic based regulator design and its complexity for partially observed automata*, in Proc. 28th Conference on Decision and Control, Tampa, FL, pp. 132–137.

- [2] H. CHO AND S. I. MARCUS, *On the supremal languages of sublanguages that arise in supervisor synthesis problems with partial observations*, Math. Control Signals Systems, 2 (1989), pp. 47–69.
- [3] R. CIESLAK, C. DESCLAUX, A. FAWAZ, AND P. VARAIYA, *Supervisory control of discrete-event processes with partial observations*, IEEE Trans. Automat. Control, 33 (1988), pp. 249–260.
- [4] J. E. HOPCROFT AND J. D. ULLMAN, *Introduction to Automata Theory, Languages, and Computation*, Addison–Wesley, Reading, MA, 1979.
- [5] R. KUMAR, V. GARG, AND S. I. MARCUS, *Language stability of dedds*, in Proc. of Internat. Conference on Mathematical Theory of Control, Bombay, India, December 1990.
- [6] F. LIN AND W. M. WONHAM, *Controllability and observability in the state-feedback control of discrete-event systems*, in Proceedings of 27th CDC, December 1988.
- [7] ———, *Decentralized control and coordination of discrete-event systems*, in Proceedings of 27th CDC, December 1988.
- [8] ———, *On observability of discrete event systems*, Inform. Sci., 44 (1988), pp. 173–198.
- [9] J. S. OSTROFF AND W. M. WONHAM, *A temporal logic approach to real time control*, in Proceedings of 24th CDC, December 1985.
- [10] C. M. ÖZVEREN, *Analysis and control of discrete event dynamic systems: A state space approach*, Ph.D. thesis, MIT, Cambridge, MA, August 1989; Laboratory for Information and Decision Systems Report LIDS-TH-1907, MIT.
- [11] C. M. ÖZVEREN AND A. S. WILLSKY, *Aggregation and multi-level control in discrete event dynamic systems*, Automatica, May 1992.
- [12] ———, *Invertibility of discrete event dynamic systems*, Math. Control Signals Systems, 1992.
- [13] ———, *Observability of discrete event dynamic systems*, IEEE Trans. Automat. Control, 35 (1990), pp. 797–806.
- [14] ———, *Output stabilizability of discrete event dynamic systems*, IEEE Trans. Automat. Control, 35 (1990), pp. 797–806.
- [15] ———, *Applications of a regulator theory for discrete event dynamic systems*, in Proceedings of IFAC Distributed Intelligence Systems Symposium, Arlington, VA, August 1991, pp. 925–935.
- [16] C. M. ÖZVEREN, A. S. WILLSKY, AND P. J. ANTSAKLIS, *Stability and stabilizability of discrete event dynamic systems*, J. Assoc. Comput. Mach., 38 (1991), pp. 730–752.
- [17] P. J. RAMADGE, *Some tractable supervisory control problems for discrete event systems modeled by buchi automata*, IEEE Trans. Automat. Control, 36 (1989), pp. 10–19.
- [18] P. J. RAMADGE AND W. M. WONHAM, *Modular feedback logic for discrete event systems*, SIAM J. Control Optim., 25 (1987), pp. 1202–1218.
- [19] ———, *Supervisory control of a class of discrete event processes*, SIAM J. Control Optim., 25 (1987), pp. 206–230.
- [20] J. N. TSITSIKLIS, *On the control of discrete event dynamical systems*, Math. C. S. S., 1989.
- [21] A. F. VAZ AND W. M. WONHAM, *On supervisor reduction in discrete event systems*, Internat. J. Control, 44 (1986), pp. 475–491.
- [22] W. M. WONHAM AND P. J. RAMADGE, *On the supremal controllable sublanguage of a given language*, SIAM J. Control Optim., 25 (1987), pp. 637–659.

INFORMATION STRUCTURES, CAUSALITY, AND NONSEQUENTIAL STOCHASTIC CONTROL I: DESIGN-INDEPENDENT PROPERTIES*

MARK S. ANDERSLAND[†] AND DEMOSTHENIS TENEKETZIS[‡]

Abstract. In control theory, the usual notion of causality—that, at all times, a system's output (action) only depends on its past and present inputs (observations)—presupposes that all inputs and outputs can be ordered, a priori, in time. In practice, many distributed systems (those subject to deadlock, for instance) are not *sequential* in this sense.

This paper explores the relationship between deadlock freeness, a less restrictive notion of causality, and the properties of a potentially *nonsequential* generic stochastic control problem formulated within the framework of Witsenhausen's intrinsic model. A property of the problem's *information structure* that is necessary and sufficient to ensure deadlock-freeness is identified and shown to be sufficient to ensure that all of the problem's control policies possess expected rewards. It is also shown, by example, that there exist stochastic control problems for which all sequential policies are suboptimal.

These results subsume Witsenhausen's "causality" condition (property C), suggest a framework for the optimization of unconstrained nonsequential stochastic control problems, and provide an intuitive design-independent characterization of the cause/effect notion of causality. The results also have game theoretic implications—they suggest, for instance, necessary and sufficient conditions for a finite game to possess an extensive form.

Key words. information structures, causality, deadlock-freeness, nonsequential stochastic control.

1. Introduction. In control theory, the usual notion of causality—that, at all times, a system's output (action) only depends on its past and present inputs (observations)—presupposes that all inputs and outputs can be ordered, a priori, in time. As it becomes increasingly attractive to decentralize the control of large systems, it has become clear that many important systems—distributed data [5], communication [13], manufacturing [11], and detection networks (Appendix A), for instance—need not be *sequential* in this sense.

The distinguishing feature of these *nonsequential* systems is the impossibility of ordering their control actions a priori, independently of the set of control laws, called the *design* (or *control policy*), that determines the actions. In the simplest case, a system's actions can be ordered a priori, given any design, but the order varies from design to design. More generally, for at least one design, the order implicitly depends on the system's uncontrolled inputs—e.g., action α may depend on action β under some circumstances while β may depend on α under others. In the worst case, for some design, and for some uncontrolled input, no "causal" ordering of the actions is possible because two or more actions are mutually dependent—e.g., action α depends on action β and vice versa. This last phenomenon, unique to nonsequential systems, is known as *deadlock*.

In this paper we explore the relationship between *deadlock-freeness*, a property that generalizes the usual notion of causality, and nonsequential stochastic control. We begin by defining deadlock-freeness (Definition 1, §3.1). Given this definition

* Received by the editors February 2, 1990; accepted for publication (in revised form) March 3, 1991. This research was supported in part by National Science Foundation grant ECS-8517708, Office of Naval Research grant N00014-87-K-0540, and a Hewlett Packard Faculty Development Award.

[†] Department of Electrical and Computer Engineering, The University of Iowa, Iowa City, Iowa 52242-1595.

[‡] Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, Michigan 48109-2122.

we consider the following question: Under what conditions is it possible to pose well-defined nonsequential stochastic control problems? This question is of interest because there exist problems for which all *sequential* designs (designs whose actions can be ordered a priori) are suboptimal (see Appendix A).

Witsenhausen's intrinsic model [19], [21] provides the framework for our results. This model, which was originally used to investigate a related causality question, encompasses all systems in which (1) the uncontrolled inputs can be viewed as an element of a measurable space (Ω, \mathcal{B}) ; (2) the number of actions to be taken is finite, say N ; (3) the k th action, $k = 1, 2, \dots, N$, can be viewed as an element of a measurable space (U^k, \mathcal{U}^k) in which the singletons are measurable; and (4) the possible designs can be viewed as N -tuples $\gamma := (\gamma^1, \gamma^2, \dots, \gamma^N)$ of $\mathcal{J}^k/\mathcal{U}^k$ -measurable functions γ^k , $k = 1, 2, \dots, N$, where the subfield \mathcal{J}^k of the product field $\mathcal{B} \otimes (\bigotimes_{i=1}^N \mathcal{U}^i)$ denotes the maximal information (knowledge) that can be used to select the k th action.

Within this framework, we identify a property of the information subfields \mathcal{J}^k , $k = 1, 2, \dots, N$, (property CI, §3.2) that is necessary and sufficient to ensure that every N -tuple γ of $\mathcal{J}^k/\mathcal{U}^k$ -measurable functions γ^k , $k = 1, 2, \dots, N$, is deadlock-free. Moreover, we show that this property is sufficient to ensure that an expected reward can be defined for every N -tuple, and consequently, that the problem of maximizing a generic system's expected reward, given a probability measure on (Ω, \mathcal{B}) , and a reward function, is well-posed. The property is *design-independent* in the sense that it holds for all designs γ .

These results subsume Witsenhausen's "causality" condition (property C) [19], suggest a framework for the recursive optimization of unconstrained nonsequential stochastic control problems [1], and provide an intuitive characterization of the cause/effect notion of causality. In essence, this characterization says that a system is causal if and only if for each tuple of uncontrolled inputs there exists an ordering of the system's actions such that no information that may be used to determine an action depends on that action or subsequent actions.

There are other approaches to the modeling of nonsequential systems. None, however, are as well suited to examining the relationship between deadlock-freeness and nonsequential control as the intrinsic model. Most game-theoretic models that accommodate nonsequentiality are variations of Kuhn's extensive form [12], a "game tree" representation that precludes deadlock by definition (cf. [19, §2]). The discrete event models that accommodate nonsequentiality are, for the most part, state transition- (e.g., [6], [16]), algebraic equation- (e.g., [9], [10], [15]), or logical calculus- (e.g., [4], [8], [14], [17]) based representations of the action sequences (traces) that a system can generate; consequently, they are incompatible with the usual control theoretic representations of uncertainty and information.

The remainder of the paper is organized as follows. In §2 we introduce Witsenhausen's intrinsic model and formulate our generic nonsequential stochastic control problem. In §3 we define properties DF (deadlock-freeness) and CI (causal implementability), and prove that property CI, a condition that is necessary and sufficient to ensure deadlock-freeness, is sufficient to ensure that unconstrained versions of the generic problem are well defined. In §4 we consider the relationship between property CI and Witsenhausen's "causality" property C. Section 5 contains our conclusions.

2. Problem formulation. To examine the relationship between deadlock-freeness and nonsequential stochastic control it is necessary to represent nonsequential systems in a framework in which each action can be viewed as depending on some system information, for instance, an observation of the system. The "conventional" con-

trol theoretic models—controlled difference, or differential equations modeling time-indexed “states” and “observations”—provide such a framework; however, they presuppose a fixed ordering of the system’s control actions. In this paper, as in [19] and [21], we relax this assumption.

2.1. Preliminaries. Consider a generic stochastic system in which the number of control actions and uncontrolled inputs are both finite (Fig. 1). From a game-theoretic perspective (cf. [18]), the control actions can be viewed as being the actions of N distinct decision-making *agents* (computers, devices, processes, etc.). Likewise the uncontrolled inputs can be viewed a single action of *nature* (chance).

To couple the agents’ actions without preordering their decisions, suppose that nature’s action $\omega := (\omega^0, \omega^1, \dots, \omega^N)$, the k th agent’s observation y^k , and the k th agent’s action u^k , take values in, respectively, the measurable spaces (Ω, \mathcal{B}) , (Y^k, \mathcal{Y}^k) , and (U^k, \mathcal{U}^k) . Let $U := \prod_{i=1}^N U^i$ and $\mathcal{U} := \bigotimes_{i=1}^N \mathcal{U}^i$; constrain the system’s k th observation to be a measurable function

$$(2.1) \quad h^k : (\Omega \times U, \mathcal{B} \otimes \mathcal{U}) \rightarrow (Y^k, \mathcal{Y}^k)$$

of the system’s *intrinsic variables*, ω and $u := (u^1, u^2, \dots, u^N)$; and constrain the k th agent’s decision policy, to be a measurable function

$$(2.2) \quad g^k : (Y^k, \mathcal{Y}^k) \rightarrow (U^k, \mathcal{U}^k)$$

of this observation.

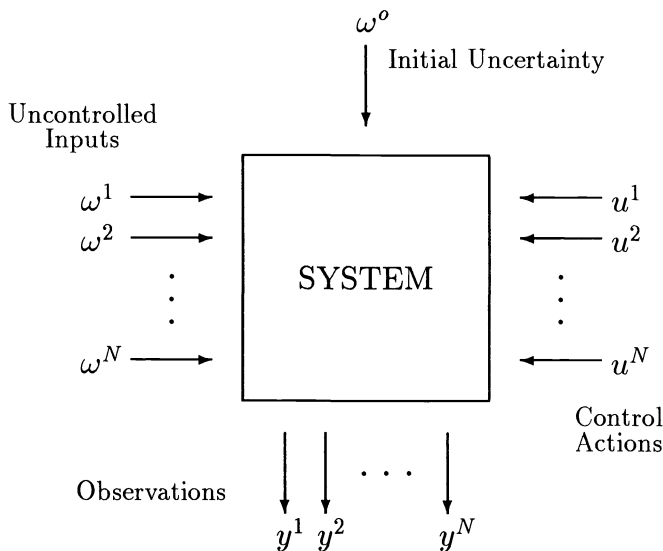


FIG. 1. A generic stochastic system.

With respect to the “conventional” discrete-time, finite horizon models of stochastic control, this representation entails no loss of generality. The system’s uncontrolled

inputs—its initial state, state and observation noises, and so on—can always be viewed as a single uncontrolled input $\omega \in \Omega$. Moreover, the k th observation—normally assumed to be a measurable function of some subset of the system's control actions, states, and random inputs—can always be viewed as a measurable function of the system's intrinsic variables.

The advantage of this representation, as opposed to the conventional models, is that as long as the superscripts on y and u are not assumed to index time, it permits interdependence among a system's control actions—e.g., given a fixed control policy $\gamma := (\gamma^1, \gamma^2, \dots, \gamma^N)$, u^j may depend on u^k (through y^j) for some ω , and vice versa for other ω . Consequently, it is possible to model nonsequential stochastic control problems, that is, problems in which a causal ordering of the control actions cannot be determined a priori because the ordering is policy, and possibly, ω -dependent.

Witsenhausen's intrinsic model [19], [21] simplifies the preceding representation. The crucial observations are (1) that the system's control actions are solely determined by the intrinsic variables (e.g., $u^k = (g^k \circ h^k)(\omega, u)$ for all $k = 1, 2, \dots, N$); and (2) that for reasonable observation functions, the k th observation, $k = 1, 2, \dots, N$, can only affect the k th control action via the information subfield it induces on the space of intrinsic variables (i.e., via $[h^k]^{-1}(\mathcal{Y}^k) \subset \mathcal{B} \otimes \mathcal{U}^1$). Accordingly, it is unnecessary to model the observations explicitly if the control agents' actions are viewed as measurable functions of the intrinsic variables.

2.2. The intrinsic model. Formally, the intrinsic model has three components:

1. An *information structure* $\mathcal{I} := \{(\Omega, \mathcal{B}), (U^k, \mathcal{U}^k), \mathcal{J}^k : 1 \leq k \leq N\}$ specifies the system's allowable decisions and distinguishable events.

(a) $N \in \mathbb{N}$ denotes the number of control actions to be taken.

(b) (Ω, \mathcal{B}) denotes the measurable space from which a random input ω is drawn.

(c) (U^k, \mathcal{U}^k) denotes the measurable space from which u^k , the k th control action, is selected. It is assumed that $\text{card}(U^k)$ is greater than one,² and that \mathcal{U}^k contains the singletons of U^k . The product space containing the N -tuple of control actions, $u := (u^1, u^2, \dots, u^N)$, is denoted by $(U, \mathcal{U}) := (\prod_{i=1}^N U^i, \otimes_{i=1}^N \mathcal{U}^i)$.³

(d) σ -field $\mathcal{J}^k \subset \mathcal{B} \otimes \mathcal{U}$ characterizes the maximal information that can be used to select the k th control action.

2. A design constraint set Γ_C constrains the set of admissible N -tuples of control laws, $\gamma := (\gamma^1, \gamma^2, \dots, \gamma^N)$, called *designs*, to a nonempty subset of $\Gamma := \prod_{i=1}^N \Gamma^i$, where Γ^k , $k = 1, 2, \dots, N$, denotes the set of all $\mathcal{J}^k/\mathcal{U}^k$ -measurable functions.

3. A probability measure P on (Ω, \mathcal{B}) specifies the mixed (randomized) decision policy to be used by nature to select ω .

Note that the intrinsic model does not exclude the possibility of an agent employing a mixed decision policy, or a policy that occasionally dictates that the agent not act. To model the mixed policy, randomizing devices can be included as factors in

¹ $[f]^{-1}$ denotes the inverse image of the function f , $[f]^{-1}(\mathcal{C}) := \{[f]^{-1}(A) : A \in \mathcal{C}\}$ denotes the set of inverse images induced by the sets in \mathcal{C} . Since inverse images preserve unions and complements, the inverse image of a σ -field is always a σ -field.

² Although the assumption $\text{card}(U^k) > 1$ was not made by Witsenhausen, it does not constitute a loss of generality. Any agent k for which $\text{card}(U^k) = 1$ has no decision to make; consequently, that agent can be deleted from the model without effect (naturally, the remaining agents' information fields—defined in 1(d)—must be adjusted to account for the k th agent's deletion—i.e., for all $j \neq k$, \mathcal{J}^j must be replaced by $\mathcal{J}^j|_{u^k}$, the u^k -section of \mathcal{J}^j).

³ $\mathcal{X} \otimes \mathcal{Y}$ denotes the product σ -field of the σ -fields \mathcal{X} and \mathcal{Y} —i.e., $\mathcal{X} \otimes \mathcal{Y} := \sigma([\pi_X]^{-1}(\mathcal{X}) \cup [\pi_Y]^{-1}(\mathcal{Y}))$, the smallest σ -field of $X \times Y$ for which the canonical projections $\pi_X(\pi_X(x, y) = x)$ and $\pi_Y(\pi_Y(x, y) = y)$ are both measurable.

(Ω, \mathcal{B}, P) , and the effects of the devices' outputs can be specified in \mathcal{J}^k . To model the occasional inaction, the agent can be allowed to make decisions that have no effect.

2.3. A generic problem. Within this framework we can formulate the following generic stochastic control problem.

- (P) Given an information structure \mathcal{I} , a design constraint set Γ_C , a probability measure P , and a bounded, nonnegative, $\mathcal{B} \otimes \mathcal{U}$ -measurable reward function V ,

Identify a design γ in Γ_C that achieves

$$\sup_{\gamma \in \Gamma_C} E_\omega[V(\omega, u_\omega^\gamma)] \text{ exactly, or within } \epsilon > 0.^4$$

Is this generic problem well defined? Since the problem may be nonsequential there are two issues: “deadlock-freeness” (Is every $\gamma \in \Gamma_C$ deadlock-free?) and “mathematical wellposedness” (Does every design $\gamma \in \Gamma_C$ possess an expected reward?).

In general, nonsequential problems of the form (P) need not be deadlock-free or well-posed. Suppose, for instance, that for some design $\gamma \in \Gamma_C$, and some random outcome $\omega \in \Omega$, the control actions

$$(2.3) \quad u^j = \gamma^j(\omega, u^1, \dots, u^k, \dots, u^N),$$

and

$$(2.4) \quad u^k = \gamma^k(\omega, u^1, \dots, u^j, \dots, u^N),$$

are interdependent. Then a deadlock arises, and consequently, the problem is not deadlock-free. Alternatively, suppose that for some design $\gamma \in \Gamma_C$, and some random outcome $\omega \in \Omega$, the *closed-loop* equations

$$(2.5) \quad u^k = \gamma^k(\omega, u^1, \dots, u^N), \quad k = 1, 2, \dots, N$$

fail to possess a unique solution

$$(2.6) \quad u_\omega^\gamma := (u_\omega^{\gamma^1}, \dots, u_\omega^{\gamma^N}).$$

Then the reward $V(\omega, u_\omega^\gamma)$ induced by ω under γ need not be unique, the expected reward $E_\omega[V(\omega, u_\omega^\gamma)]$ need not exist, and consequently, the problem need not be well posed.

The primary objective of this paper is to identify conditions sufficient to ensure that problem (P) is deadlock-free and well-posed. Since there exist problems of the form (P) for which some, but not all, nontrivial designs are deadlock-free and possess expected rewards (Appendix A)—two classes of conditions can be considered: conditions based on the problem's *design-independent* properties (properties that hold for *all* $\gamma \in \Gamma$), and conditions based on the problem's *design-dependent* properties (properties that may only hold for *specific* designs $\gamma \in \Gamma$). In this paper, conditions based on the problem's *design-independent* properties are explored. Conditions based on the problem's *design-dependent* properties are introduced in a companion paper [3].

⁴ The notation u_ω^γ indicates that u depends on ω through γ (see §3.1).

3. Design-independent conditions. In this section, necessary and sufficient conditions for problem (P) to be well-posed and deadlock-free are developed under the assumption that the problem's design set is unconstrained (i.e., $\Gamma_C = \Gamma$). The conditions are design-independent in the sense that they are solely based on properties of the problem's information structure \mathcal{I} .

3.1. Properties DF, S, and SM. To ensure that problem (P) is deadlock-free it suffices to require that its information structure \mathcal{I} possess *property* DF (deadlock-freeness).

DEFINITION 1. An information structure \mathcal{I} possesses *property* DF (*deadlock-freeness*) if for each $\gamma \in \Gamma$, and for every $\omega \in \Omega$, there exists an ordering of γ 's N control laws, say $\gamma^{s_1(\omega)}, \gamma^{s_2(\omega)}, \dots, \gamma^{s_N(\omega)}$, such that no control action $u^{s_n(\omega)}$, $n = 1, 2, \dots, N$, depends on itself or the control actions that follow.

Note that the ordering in Definition 1 may depend on the design $\gamma \in \Gamma$ and the random input $\omega \in \Omega$. For instance, for some $\gamma \in \Gamma$ a triggering random event may determine the identity of the initial control action (see Appendix A).

When \mathcal{I} possesses property DF, for each $\gamma \in \Gamma$ and for all $\omega \in \Omega$, γ is deadlock-free in the sense that, given ω , $u^{s_1(\omega)}$ can be determined; given ω and $u^{s_1(\omega)}$, $u^{s_2(\omega)}$ can be determined; given ω , $u^{s_1(\omega)}$ and $u^{s_2(\omega)}$, $u^{s_3(\omega)}$ can be determined; and so on. Hence, property DF generalizes the usual notion of causality in the sense that it does not presuppose that the actions' order is fixed.

To ensure that problem (P) is well-posed, it suffices to require (i) that for each $\gamma \in \Gamma$ and every $\omega \in \Omega$ there exist a unique $u := (u^1, u^2, \dots, u^N) \in U$ satisfying the system of equations

$$(3.1) \quad u^k = \gamma^k(\omega, u), \quad k = 1, 2, \dots, N,$$

and (ii) that each of the *solution maps* $\Sigma^\gamma : \Omega \rightarrow U$ induced via these solutions (i.e., $\Sigma^\gamma(\omega) = u_\omega^\gamma$ where $u_\omega^\gamma = \gamma(\omega, u_\omega^\gamma)$) be \mathcal{B}/\mathcal{U} -measurable. Then, for each $\gamma \in \Gamma$, $V(\cdot, \Sigma^\gamma(\cdot))$ is \mathcal{B} -measurable, and consequently, $E_\omega[V(\omega, \Sigma^\gamma(\omega))]$ exists. Systems that satisfy (i) are said to possess *property* S (solvability) while systems that satisfy (ii) are said to possess *property* SM (solvability/measurability) [19]. In fact, property S often implies property SM [2].

3.2. Property CI. Property SM holds when, for each $\gamma \in \Gamma$, and each uncontrolled input $\omega \in \Omega$, every agent's action is uniquely determined and the actions' ω -dependence is \mathcal{B} -measurable. Since property SM does not rule out the possibility that, for some $\omega \in \Omega$, agent N 's information depends on agent 1's action, and for all $k = 1, 2, \dots, N - 1$, agent k 's information depends on agent $k + 1$'s action, property SM is not sufficient to ensure property DF (cf. [19], Thm. 2). That is, although property SM holds, for some $\omega \in \Omega$, every agent's information may depend on every other agents' actions, and consequently, for that ω , no agent can act without precognition.

Property DF suggests that such deadlocks cannot arise if for each $\omega \in \Omega$, the agents can be ordered such that each agent's information only depends on ω and its predecessors' actions. To formalize this observation it is convenient to adopt the notation in [19]. For all $k = 1, 2, \dots, N$, define S_k to be the set of all k -agent orderings—i.e., all injections of $\{1, 2, \dots, k\}$ into $\{1, 2, \dots, N\}$. For all $j = 0, 1, \dots, N$, and $k = j, j + 1, \dots, N$, let $T_j^k : S_k \rightarrow S_j$ denote a truncation map that returns the ordering of the first j agents of a k -agent ordering—i.e., T_j^k restricts $s \in S_k$ to the domain $\{1, 2, \dots, j\}$ or to \emptyset when $j = 0$. Finally, for all $s := (s_1, s_2, \dots, s_k) \in S_k$, and

$k = 1, 2, \dots, N$, define \mathcal{P}_s to be the projection of $\Omega \times U$ onto $\Omega \times (\prod_{i=1}^k U^{s_i})$ —i.e.,

$$(3.2) \quad \mathcal{P}_s(\omega, u) := (\omega, u^{s_1}, u^{s_2}, \dots, u^{s_k}), \quad \mathcal{P}_\emptyset(\omega, u) := (\omega).$$

Then, we can characterize deadlock-freeness as follows.

DEFINITION 2. An information structure \mathcal{I} possesses *property CI (causal implementability)* when there exists at least one map $\psi : \Omega \times U \rightarrow S_N$ such that for all $k = 1, 2, \dots, N$, and $(\omega, u) \in \Omega \times U$,

$$(3.3) \quad \mathcal{J}^{s_k} \cap [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega, u)) \subset \{\emptyset, [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega, u))\}$$

when $s := (s_1, s_2, \dots, s_N) = \psi(\omega, u)$.

ψ is a function that maps every intrinsic outcome $(\omega, u) \in \Omega \times U$ into an N -agent ordering.

$$(3.4) \quad [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega, u)) = [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\omega, u^{s_1}, \dots, u^{s_{k-1}})$$

is the cylinder set induced on $\Omega \times U$, when the intrinsic outcome is (ω, u) , by the actions of nature and the first $k - 1$ agents in $s := (s_1, s_2, \dots, s_N) = \psi(\omega, u)$. Since

$$(3.5) \quad \mathcal{J}^{s_k} \cap [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega, u))$$

denotes the *trace* of the s_k th agent's information field on this cylinder set (i.e., $\mathcal{J}^{s_k} \cap C := \{A \cap C : A \in \mathcal{J}^{s_k}\}$), (3.3) constrains the cylinder set to be a subset of all events containing (ω, u) in the s_k th agent's information field \mathcal{J}^{s_k} —i.e., no event in \mathcal{J}^{s_k} containing (ω, u) may depend on $u^{s_k}, u^{s_{k+1}}, \dots$, or u^{s_N} . Accordingly, property CI ensures that for all outcomes $(\omega, u) \in \Omega \times U$, there exists an order $s := (s_1, s_2, \dots, s_N) = \psi(\omega, u)$ such that, for all $k = 1, 2, \dots, N$, the s_k th agent's information, at the point (ω, u) , only depends on the actions of nature and its predecessors in s .

Property CI implies property SM and is a necessary and sufficient condition for all $\gamma \in \Gamma$ to be deadlock-free. Theorem 1 states this formally.

THEOREM 1. Let \mathcal{I} be an arbitrary information structure, then

- (i) \mathcal{I} possesses property SM if \mathcal{I} possesses property CI, and
- (ii) \mathcal{I} possesses property DF if and only if \mathcal{I} possesses property CI.

Proof. See Appendix B. \square

Theorem 1 ensures that problem (P) is *well defined* (deadlock-free and well-posed) when it satisfies property CI. Its proof hinges on the following observation. When ψ is an order function such that \mathcal{I} possesses property CI, for arbitrary but fixed $(\omega, u) \in \Omega \times U$, and $k = 1, 2, \dots, N$, (3.3) and the fact that \mathcal{U}^k contains the singletons of U^k imply that, at the point (ω, u) , all $\mathcal{J}^{s_k}/\mathcal{U}^{s_k}$ -measurable functions $\gamma^{s_k} \in \Gamma^{s_k}$, $s := (s_1, s_2, \dots, s_N) = \psi(\omega, u)$, do not depend on the components s_k, s_{k+1}, \dots , and s_N of u . This suggests that, for fixed $\gamma \in \Gamma$, a unique \mathcal{B} -measurable solution $\Sigma^\gamma : \Omega \rightarrow U$ to the closed-loop equation $u = \gamma(\omega, u)$ can be obtained by the following recursion. Fix $\omega \in \Omega$, let $r \in U$ be an arbitrary reference element, let $L^\gamma : \Omega \times U \rightarrow \Omega \times U$ be defined as

$$(3.6) \quad L^\gamma(\omega, r) := (\omega, \gamma(\omega, r)),$$

and let $L_k^\gamma : \Omega \times U \rightarrow \Omega \times U$ be a k -fold composition of L^γ —i.e.,

$$(3.7) \quad L_k^\gamma(\omega, r) := (\underbrace{L^\gamma \circ \dots \circ L^\gamma}_{k \text{ times}})(\omega, r).$$

1. After one iteration, the components of $L_1^\gamma(\omega, r)$ corresponding to agents whose information, at the point (ω, r) , does not depend on r , become invariant to subsequent iterations. By property CI, the set $\mathcal{A}_1(\omega) \subset \{1, 2, \dots, N\}$ indexing (by agent) these components is nonempty since at least agent $(\psi(\omega, r))_1$'s information does not depend on r .

2. After two iterations, the components of $L_2^\gamma(\omega, r)$ corresponding to agents in $\{1, 2, \dots, N\} \setminus \mathcal{A}_1(\omega)$ whose information, at the point $L_1^\gamma(\omega, r)$, does not depend on the components of agents in $\{1, 2, \dots, N\} \setminus \mathcal{A}_1(\omega)$, become invariant to subsequent iterations.⁵ By property CI, the set $\mathcal{A}_2(\omega)$ indexing (by agent) these components is nonempty when $\text{card}(\mathcal{A}_1(\omega)) < N$ since at least agent $(\psi(L_1^\gamma(\omega, r)))_j$'s information,

$$(3.8) \quad j = \min\{m \in \{1, 2, \dots, N\} : (\psi(L_1^\gamma(\omega, r)))_m \notin \mathcal{A}_1(\omega)\},$$

does not depend on the components of agents in $\{1, 2, \dots, N\} \setminus \mathcal{A}_1(\omega)$.

⋮

k. After k iterations, the components of $L_k^\gamma(\omega, r)$ corresponding to agents in $\{1, 2, \dots, N\} \setminus \bigcup_{i=1}^{k-1} \mathcal{A}_i(\omega)$ whose information, at the point $L_{k-1}^\gamma(\omega, r)$, does not depend on the components of agents in $\{1, 2, \dots, N\} \setminus \bigcup_{i=1}^{k-1} \mathcal{A}_i(\omega)$, become invariant to subsequent iterations. By property CI, the set $\mathcal{A}_k(\omega)$ indexing (by agent) these components is nonempty when $\text{card}(\bigcup_{i=1}^{k-1} \mathcal{A}_i(\omega)) < N$ since at least agent $(\psi(L_{k-1}^\gamma(\omega, r)))_j$'s information,

$$(3.9) \quad j = \min\left\{m \in \{1, 2, \dots, N\} : (\psi(L_{k-1}^\gamma(\omega, r)))_m \notin \bigcup_{i=1}^{k-1} \mathcal{A}_i(\omega)\right\}$$

does not depend on the components of agents in $\{1, 2, \dots, N\} \setminus \bigcup_{i=1}^{k-1} \mathcal{A}_i(\omega)$.

⋮

and so on.

Since property CI ensures that, until all agents' components are invariant, at least one new component becomes invariant after every iteration, the recursive procedure must converge in, at most, N iterations—i.e., the unique solution to the closed-loop equation $u = \gamma(\omega, u)$ is $\pi_U(L_N^\gamma(\omega, r))$, where π_U denotes the canonical projection of $\Omega \times U$ onto U ($\pi_U(\omega, u) = u$) and $r \in U$ is an arbitrary “seed” that starts the recursive solution process. Since π_Ω , π_U , and γ are, respectively, $\mathcal{B} \otimes \mathcal{U}/\mathcal{B}$ -, $\mathcal{B} \otimes \mathcal{U}/\mathcal{U}$ -, and $\mathcal{B} \otimes \mathcal{U}/\mathcal{U}$ -measurable, L^γ , and by composition, L_k^γ and $\pi_U \circ L_N^\gamma$, are, respectively, $\mathcal{B} \otimes \mathcal{U}/\mathcal{B} \otimes \mathcal{U}$ -, $\mathcal{B} \otimes \mathcal{U}/\mathcal{B} \otimes \mathcal{U}$ -, and $\mathcal{B} \otimes \mathcal{U}/\mathcal{U}$ -measurable. It follows, since all u -sections of $\mathcal{B} \otimes \mathcal{U}/\mathcal{U}$ -measurable functions are \mathcal{B}/\mathcal{U} -measurable, that the induced solution map $\Sigma^\gamma = \pi_U \circ L_N^\gamma|_r$ is necessarily \mathcal{B}/\mathcal{U} -measurable.

The above recursion has an obvious physical interpretation. For fixed $\gamma \in \Gamma$ and $\omega \in \Omega$, suppose that we conduct the following thought experiment: decouple the agents and record in succession, $C_1(\omega)$, the indices of those agents that act given ω

⁵ For sets $A, B \subset X$, $A \setminus B := \{x \in A : x \notin B\}$.

alone; $C_2(\omega)$, the indices of those agents that act given ω and the actions of agents in $C_1(\omega)$; $C_3(\omega)$, the indices of those agents that act given ω and the actions of agents $C_1(\omega) \cup C_2(\omega)$; and so on. Clearly, $\mathcal{A}_k(\omega) = C_k(\omega)$ for all $k = 1, 2, \dots, N$. Accordingly, if for all k we ignore all components of $\pi_U(L_k^\gamma(\omega, r))$ but those corresponding to the agents indexed in $\mathcal{A}_k(\omega)$, the preceding recursion outlines the partial ordering of agent actions that a passive observer would record, given ω , if the design γ were implemented in a “maximally” concurrent fashion.

Although the preceding recursion implicitly demonstrates that property CI implies property DF, it is far easier to establish sufficiency by a direct appeal to property CI. For all $(\omega, u) \in \Omega \times U$ and $k = 1, 2, \dots, N$, property CI implies that at the point (ω, u) , all $\mathcal{I}^{s_k}/\mathcal{U}^{s_k}$ -measurable functions $\gamma^{s_k} \in \Gamma^{s_k}$, $s := (s_1, s_2, \dots, s_N) = \psi(\omega, u)$, do not depend on the components s_k, s_{k+1}, \dots , and s_N of u . Consequently, no agent’s information depends on its own action or the actions of its successors—i.e., the system must be deadlock-free.

The fact that some design $\gamma \in \Gamma$ must deadlock when property CI fails to hold is also a direct consequence of property CI’s definition. When property CI fails, for some outcome $(\omega, u) \in \Omega \times U$ and for all N -agent orderings $s \in S_N$, (3.3) fails for at least one $k \in \{1, 2, \dots, N\}$. Since there are at most $N \text{card}(S_N) = N(N!)$ k, s combinations for which (3.3) can fail, and since all agents may take at least two distinct actions, it is always possible to construct a design γ that possesses all of the interdependencies that cause (3.3) to fail—i.e., a design γ such that for all $s \in S_N$, when the s_k th agent’s information depends on the actions of its successors in s , $\gamma^{s_k}(\omega, u)$ depends on the s_k th agent’s successors’ components of u . Accordingly, it is always possible to construct a design that deadlocks.

4. Property CI’s relationship to property C. Witsenhausen was the first to develop conditions sufficient to ensure a system’s deadlock-freeness (he termed it “causality”). Specifically, he introduced the following property.

DEFINITION 3 ([19]). An information structure \mathcal{I} possesses *property C* (*causality*) when there exists at least one map $\psi : \Omega \times U \rightarrow S_N$ such that for all $s := (s_1, s_2, \dots, s_k) \in S_k, k = 1, 2, \dots, N$,

$$(4.1) \quad \mathcal{I}^{s_k} \cap [T_k^N \circ \psi]^{-1}(s) \subset \mathcal{F}(T_{k-1}^k(s)),$$

where $\mathcal{F}(s)$ denotes the cylindrical extension of $\mathcal{B} \otimes (\bigotimes_{i=1}^k \mathcal{U}^{s_i})$ to $\Omega \times U$ for all $s \in S_k, k = 1, 2, \dots, N$.⁶

He then proved the following theorem (DF is our terminology).

THEOREM 2 ([19]). *Let \mathcal{I} be an arbitrary information structure; then*

- (i) *\mathcal{I} possesses property SM if \mathcal{I} possesses property C, and*
- (ii) *\mathcal{I} possesses property DF if \mathcal{I} possesses property C.*

Proof. See [19, §§6 and 7]. \square

Since property C implies property DF (Theorem 2(ii)), and since property DF implies property CI (Theorem 1(ii)), the following is clear.

COROLLARY 1. *Property C implies Property CI.*

Proof. See Appendix C for a direct proof. \square

This corollary suggests that the ψ/γ -dependent umpire recursion that Witsenhausen used to prove Theorem 2 ([19, §7]), is not fundamental—i.e., to prove Theorem 2, it suffices to compose $\gamma|_\omega$ with itself N times (i.e., to form $\pi_U \circ L_N^\gamma|_\omega$) as

⁶ Here, in contrast to [19], ψ is a mapping from $\Omega \times U$ to S_N and $\mathcal{F}(\emptyset)$ is the cylindrical extension of \mathcal{B} to $\Omega \times U$ (see [21]).

described in §3. The corollary also raises the following question: Are properties C and CI equivalent? Equivalence would imply, by Theorem 1(ii), that \mathcal{I} 's possession of property C is both a necessary and sufficient condition for deadlock-freeness. Non-equivalence would imply that property C is, in general, only a sufficient condition for deadlock-freeness.

When $N \leq 2$, properties C and CI are always equivalent. Corollary 2 states this formally.

COROLLARY 2. *Property CI implies property C when $N \leq 2$.*

Proof. By Theorem 1(i) property CI implies property SM which, in turn, implies property S. The corollary follows since property S implies property C when $N \leq 2$ ([19, Thm. 2]). \square

When $N > 2$, it is not known (in general) whether property CI implies property C; the implication, however, holds in at least two important special cases (Thms. 3 and 4).

DEFINITION 4. An information structure \mathcal{I} is said to be *sequential* when property CI holds for some constant order function ψ .

THEOREM 3. *All constant order functions ψ such that \mathcal{I} possesses property CI are order functions such that \mathcal{I} possesses property C; consequently, property CI implies property C when \mathcal{I} is sequential.*

Proof. See Appendix D. \square

Note that an unconstrained problem of the form (P) is sequential (in the sense discussed in §1) if and only if its information structure is sequential. Witsenhausen defines an information structure to be sequential when property C holds with a constant order function ψ ([20, §3]). Accordingly, Theorem 3 ensures that, as far as unconstrained problems of the form (P) are concerned, sequentiality, as defined in this paper, is equivalent to Witsenhausen's sequentiality.

When \mathcal{I} is nonsequential, even if \mathcal{I} possesses property C, order functions for which \mathcal{I} possesses property CI need not be order functions for which \mathcal{I} possesses property C.

Example 1. Consider a nonsequential information structure \mathcal{I} of the following form:

$$\begin{aligned}
 N &= 3, \\
 \Omega &= U^1 = U^2 = U^3 = \{0, 1\}, \\
 \mathcal{B} &= \mathcal{U}^1 = \mathcal{U}^2 = \mathcal{U}^3 = \{\emptyset, \{0\}, \{1\}, \{0, 1\}\}, \\
 (4.2) \quad \mathcal{J}^1 &= \{\emptyset, \{(\omega, u) : \omega u^2 = 0\}, \{(\omega, u) : \omega u^2 = 1\}, \Omega \times U\}, \\
 \mathcal{J}^2 &= \{\emptyset, \{(\omega, u) : \bar{\omega} u^1 = 0\}, \{(\omega, u) : \bar{\omega} u^1 = 1\}, \Omega \times U\}, \quad \text{and} \\
 \mathcal{J}^3 &= \{\emptyset, \{(\omega, u) : \omega = 0\}, \{(\omega, u) : \omega = 1\}, \Omega \times U\}.^7
 \end{aligned}$$

Although

$$(4.3) \quad \bar{\psi}(\omega, u^1, u^2, u^3) = \begin{cases} (1, 2, 3) & \text{when } \omega = 0 \\ (2, 1, 3) & \text{else} \end{cases}$$

⁷ \bar{x} denotes the binary complement of $x \in \{0, 1\}$ —i.e., $\bar{x} := 1 - x$.

is an order function such that \mathcal{I} possesses properties CI and C,

$$(4.4) \quad \psi(\omega, u^1, u^2, u^3) = \begin{cases} (1, 2, 3) & \text{when } \omega = 0 \\ (3, 2, 1) & \text{when } \omega u^3 = 1 \\ (2, 1, 3) & \text{else} \end{cases}$$

is an order function such that \mathcal{I} possesses property CI, but not property C ((3.3) fails when $k = 1$ and $s = 3 \in S_1$, for instance, since $[T_1^3 \circ \psi]^{-1}(3) = \{(\omega, u) : \omega u^3 = 1\} \notin \mathcal{F}(\emptyset) = \mathcal{B} \otimes \{\emptyset, U\}$).

The fact that there exist nonsequential information structures \mathcal{I} , and order functions ψ , such that \mathcal{I} possesses property CI, but not property C, implies that general proofs that property CI implies property C (if such exist) must be constructive—i.e., given a ψ such that \mathcal{I} possesses property CI, but not property C, we must be able to construct a new order function $\hat{\psi}$ (obviously distinct from ψ), such that \mathcal{I} possesses property C.

Given the generality of the intrinsic model, such constructions are, at best, tedious. Consider Example 1. By simple combinatorial arguments it can be shown that 3^{16} of the 6^{16} possible order functions for \mathcal{I} are order functions for which \mathcal{I} possesses property CI. Of these 3^{16} order functions, only 25 are order functions for which \mathcal{I} possesses property C.⁸ Any proof that property CI implies property C, under conditions satisfied by the Example 1's information structure, must produce, as a byproduct, a construction that maps every one of the 3^{16} order functions for which \mathcal{I} possesses property CI to one of the 25 order functions for which \mathcal{I} possesses property C.

One such construction (Appendix E, (E.6)-(E.11)) can be used to prove the following theorem.

THEOREM 4. *Property CI implies property C when Ω , and U^k , $k = 1, 2, \dots, N$, are countable sets, and \mathcal{B} contains the singletons of Ω .*

Proof. See Appendix E. \square

Since the success of this construction hinges on the fact that for all $s \in S_k$, $k = 0, 1, \dots, N$, $\mathcal{F}(s)$ is the cylindrical extension of the power set of $\Omega \times (\prod_{i=1}^k U^{s_k})$ (a property that only holds under the conditions of the theorem), other constructions must be developed to establish that property CI implies property C under more general conditions.

5. Conclusions. In this paper we have introduced conditions necessary and sufficient to ensure that a generic stochastic system, represented within the framework of Witsenhausen's intrinsic model, is deadlock-free. The main results concern the fact that \mathcal{I} 's possession of property CI is

(1) A necessary and sufficient condition for all $\gamma \in \Gamma$ to be deadlock-free (Theorem 1(ii)); and

(2) A sufficient condition to ensure the existence, for all $\gamma \in \Gamma$, of a unique \mathcal{B}/\mathcal{U} -measurable function Σ^γ mapping all $w \in \Omega$ into unique solutions u_w^γ of the closed-loop equation $\gamma(w, u) = u$ (Theorem 1(i)).

⁸ There are $(3!)^{16} = 6^{16}$ possible order functions $\psi : \Omega \times U \rightarrow S_3$ since $\text{card}(\Omega \times U) = 16$ and $\text{card}(S_3) = 3!$. Only 3^{16} of these satisfy the conditions of property CI since u^1 must precede u^2 when $w = 0$ and vice versa when $w = 1$ (for each (w, u) this rules out half of the $3!$ possible orders). Only 5^2 of the order functions satisfy the conditions of property C since $[T_1^3 \circ \psi]^{-1}(s)$ must be $\mathcal{F}(\emptyset)$ -measurable and $[T_2^3 \circ \psi]^{-1}(s)$ must be $\mathcal{F}(T_1^2(s))$ -measurable for all $s \in S_2$ (when $w = 0$, only $\psi|_{w=0} = (3, 1, 2)$ and $\psi|_{w=0, u^1} \in \{(1, 2, 3), (1, 3, 2)\}$ are acceptable; when $w = 1$ only $\psi|_{w=1} = (3, 2, 1)$ and $\psi|_{w=1, u^2} \in \{(2, 1, 3), (2, 3, 1)\}$ are acceptable).

These results subsume the principal result in [19 Theorem 1], and provide a necessary and sufficient condition for unconstrained stochastic control problems of the form (P) to be well-posed and deadlock-free.

The remaining results establish

(3) That \mathcal{I} 's possession of property CI ensures, for all $\gamma \in \Gamma$, that the function Σ^γ can be determined recursively, starting from an arbitrary "seed" $r \in U$, by composing $\gamma|_w$ with itself N times (see the discussion following Theorem 1);

(4) That property CI implies property C in a least three special cases (Corollary 2, Theorem 3, and Theorem 4); and

(5) That any general proof that property CI implies property C (i.e., that property C is a necessary condition for causality) must be constructive (see Example 1 and the discussion that follows).

Note that nowhere in the paper was any property of the reward—let alone the implicit assumption that agents cooperate to maximize this reward—ever used to construct a definition or derive a result; consequently, the results of this paper apply to games as well as controlled systems. For instance, by Theorem 1, a game involving a finite number of decisions chosen from decision spaces satisfying the constraints imposed by the intrinsic model, has an *extensive form* (i.e., a "game tree" representation, see [12]) if and only if its information structure possesses property CI.

Appendix A.

A. Decentralized detection: An example. This appendix concerns a decentralized detection network in which the optimal control policies must make explicit use of the fact that the network's control actions can be nonsequential. By example, it is shown that the introduction of nonsequentiality into the network can, under some circumstances, give rise to deadlocks, and under other circumstances, improve network performance.

A.1. The problem. Consider the problem of designing a simple decentralized detection network (Fig. A.1) consisting of two detectors, D1 and D2.

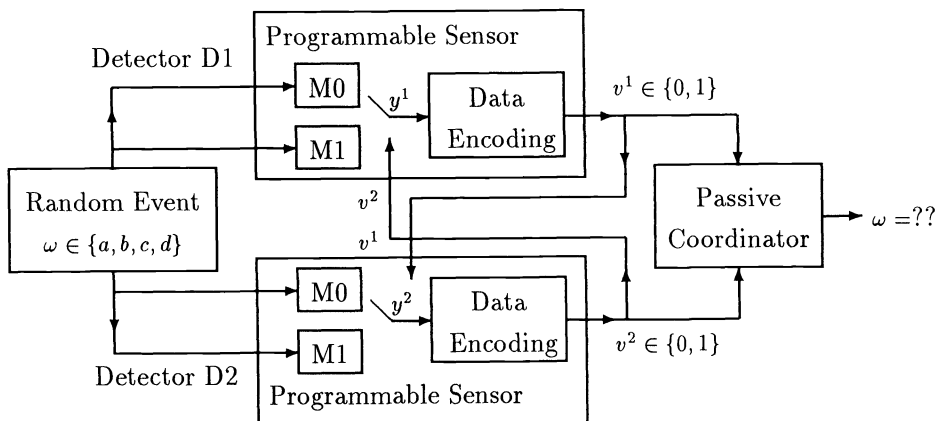


FIG. A.1. A simple decentralized detection network.

A.1.1. Observations. Each detector is permitted to make a single noisy observation, $y^k \in \{A, B\}$, $k = 1, 2$, of a random event $\omega \in \{a, b, c, d\}$ using one of two distinct configurations (Fig. A.2) of a programmable sensor possessing two operational modes $m^k \in \{0, 1\}$. Formally, for $k = 1, 2$,

$$(A.1) \quad y^k = h_{\alpha^k}(\omega, m^k),$$

where $\alpha^k \in \{1, 2\}$ indexes detector D_k 's sensor configuration.

Configuration 1:

$$h_1(\omega, m^k) = \begin{cases} A & (\omega, m^k) \in \{(a, 1), (d, 1)\} \\ A & \omega = c \\ B & (\omega, m^k) \in \{(a, 0), (d, 0)\} \\ B & \omega = b \end{cases}$$

Configuration 2:

$$h_2(\omega, m^k) = \begin{cases} A & (\omega, m^k) \in \{(b, 1), (c, 1)\} \\ A & \omega = a \\ B & (\omega, m^k) \in \{(b, 0), (c, 0)\} \\ B & \omega = d \end{cases}$$

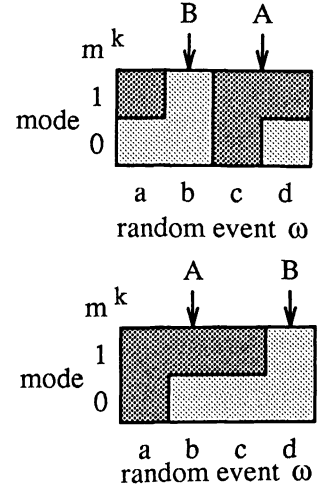


FIG. A.2. Available sensor configurations.

A.1.2. Data encoding. Once a detector has made its observation, it transmits a one bit summary, $v^k \in \{0, 1\}$, to a passive coordinator over a noiseless channel. Formally, for $k = 1, 2$,

$$(A.2) \quad v^k = g^k(y^k),$$

where g^k can be any function mapping $\{A, B\}$ to $\{0, 1\}$.

A.1.3. Sensor programming. Each detector can monitor the other's transmissions; accordingly, either may elect to program its sensor (i.e., set $m^k = 0$ or 1) based on the other's summary. Formally, for $k = 1, 2$,

$$(A.3) \quad m^k = f^k(v^{\bar{k}}),$$

where f^k can be any function mapping $\{0, 1\}$ to $\{0, 1\}$. When f^k is a constant function, the sensor programming is *static*—i.e., the mode in which detector D_k 's sensor is operated is determined a priori. When f^k is not a constant function, the sensor programming is *dynamic*—i.e., the mode in which detector D_k 's sensor is operated may depend on detector $D_{\bar{k}}$'s one bit summary (\bar{k} denotes the binary complement of $k \in \{0, 1\}$). It is the possibility that both detectors' sensors may be programmed dynamically that makes this decentralized detection network nonsequential—i.e., when neither f^1 nor f^2 is constant, the detectors' data summaries may be interdependent.

A.1.4. Passive coordinator. The passive coordinator, given the detectors' data summaries v^1 and v^2 , attempts to correctly detect (identify) the uncertain outcome $\omega \in \{a, b, c, d\}$. Formally, the coordinator generates an estimate of ω ,

$$(A.4) \quad \hat{\omega} = \eta(v^1, v^2),$$

using any function η mapping $\{0, 1\} \times \{0, 1\}$ to $\{a, b, c, d\}$.

A.1.5. Objective. Given a probability distribution for ω , the objective is to select an estimation policy for the passive observer, and sensor configurations, sensor programming policies, and data encoding policies for the detectors, that collectively maximize the probability that the coordinator can correctly identify ω . Formally, the objective is to

$$(A.5) \quad \begin{aligned} &\text{Identify a } \text{design}(\alpha^1, \alpha^2, f^1, f^2, g^1, g^2, h^1, h^2, \eta) \\ &\text{that achieves } \max_{\alpha^k, f^k, g^k, h^k, \eta} P\{\omega \in \Omega : \omega = \hat{\omega}\} \text{ exactly.}^9 \end{aligned}$$

A.2. Deadlock. Clearly the preceding detection network is susceptible to deadlock. Suppose, for instance,

- (1) That both detectors' sensors are in configuration 1 (i.e., $\alpha^1 = \alpha^2 = 1$),
- (2) That each detector programs its sensor based on the other's data summary (e.g., $m^1 = u^2$, and $m^2 = u^1$), and
- (3) That neither detector's data encoding policy is constant.

Then, when $\omega \in \{a, d\}$, detector D1's observation depends on detector D2's data summary and detector D2's observation depends on detector D1's data summary; consequently, neither detector can generate a data summary without precognition—i.e., the network is deadlocked.

A.3. A solution. Although the possibility of deadlock can be completely eliminated by constraining the network's design to be sequential (i.e., by prohibiting at least one detector from programming its sensor based on the other detector's data summary and thereby eliminating the possibility of nonsequentiality), this “fix” ignores the possibility that nonsequentiality may improve network performance. In fact,

- (1) There exists a deadlock-free nonsequential design that enables the coordinator to correctly identify, with certainty, all uncertain outcomes $\omega \in \{a, b, c, d\}$, and
- (2) No sequential design permits the coordinator to correctly identify, with certainty, more than two of the four uncertain outcomes $\omega \in \{a, b, c, d\}$.

In other words, in this case, *optimal network performance can only be achieved by exploiting the nonsequentiality of the network.*

A.3.1. An optimal nonsequential design. Consider, for instance, the following design:

⁹ Note that, although it is tedious, it is not difficult to transform this problem into an unconstrained problem of the form (P) (§2.3). By setting $\omega = \omega \in \Omega := \{a, b, c, d\}$, $u^1 = \alpha^1 \in U^1 := \{1, 2\}$, $u^2 = \alpha^2 \in U^2 := \{1, 2\}$, $u^3 = m^1 \in U^3 := \{0, 1\}$, $u^4 = m^2 \in U^4 := \{0, 1\}$, $u^5 = v^1 \in U^5 := \{0, 1\}$, $u^6 = v^2 \in U^6 := \{0, 1\}$, $u^7 = \hat{\omega} \in U^7 := \{a, b, c, d\}$; by translating the informational constraints imposed (by the original problem formulation) into constraints on the information subfields $\mathcal{J}^k, k = 1, 2, \dots, 7$, of $2^{\Omega \times U}$ (e.g., $\mathcal{J}^1 = \{\emptyset, \Omega \times U\}$ since $u^1 = \alpha^1$ must be a constant, $\mathcal{J}^3 = \{\emptyset, \Omega\} \otimes (\bigotimes_{i=1}^5 \{\emptyset, U^i\}) \otimes 2^{U^6} \otimes \{\emptyset, U^7\}$ since $u^3 = m^1$ can only depend on $u^6 = v^2$, and so on); and by setting $V(\omega, u) = I_{\{\omega = u^7\}}$ (the indicator of the event $\{\omega = u^7\}$), one can transform the original problem into an unconstrained 7-agent problem in which the first two agents' decisions determine the detectors' sensor configurations, the third and fourth agents' decisions correspond to the detectors' sensor programming decisions, the fifth and sixth agents' decisions correspond to the detectors' data summaries, the seventh agent's decision corresponds to the passive coordinator's estimate.

Detector D1:

$$(A.6) \quad \begin{aligned} m^1 &= f^1(v^2) = v^2 && \text{(D1's mode = D2's data summary),} \\ y^1 &= h_1(\omega, m^1) && \text{(D1's sensor in configuration 1), and} \\ v^1 &= g^1(y^1) = \begin{cases} 1 & y^1 = A \\ 0 & y^1 = B \end{cases} && \text{(D1's data summary).} \end{aligned}$$

Detector D2:

$$(A.7) \quad \begin{aligned} m^2 &= f^2(v^1) = v^1 && \text{(D2's mode = D1's data summary),} \\ y^2 &= h_2(\omega, m^2) && \text{(D2's sensor in configuration 2), and} \\ v^2 &= g^2(y^2) = \begin{cases} 0 & y^2 = A \\ 1 & y^2 = B \end{cases} && \text{(D2's data summary).} \end{aligned}$$

Passive Coordinator:

$$(A.8) \quad \hat{\omega} = \eta(v^1, v^2) = \begin{cases} a & (v^1, v^2) = (0, 0) \\ b & (v^1, v^2) = (0, 1) \\ c & (v^1, v^2) = (1, 0) \\ d & (v^1, v^2) = (1, 1). \end{cases}$$

It is not difficult to verify that:

- When $\omega = a$, D2 transmits $v^2 = 0$ first,
D1 transmits $v^1 = 0$ second, and
the passive coordinator sets $\hat{\omega} = a$;
- When $\omega = b$, D1 transmits $v^1 = 0$ first,
D2 transmits $v^2 = 1$ second, and
the passive coordinator sets $\hat{\omega} = b$;
- When $\omega = c$, D1 transmits $v^1 = 1$ first,
D2 transmits $v^2 = 0$ second, and
the passive coordinator sets $\hat{\omega} = c$;
- When $\omega = d$, D2 transmits $v^2 = 1$ first,
D1 transmits $v^1 = 1$ second, and
the passive coordinator sets $\hat{\omega} = d$.









Since the order in which the detectors transmit their data summaries cannot be pre-specified, this design is nonsequential. Since both detectors can transmit data summaries, without precognition, for all $\omega \in \{a, b, c, d\}$, the design is also deadlock-free. Finally, since the passive coordinator can correctly identify, with certainty, all uncertain events $\omega \in \{a, b, c, d\}$, the design is optimal.

A.3.2. No sequential design is optimal. Since the selection of an estimation policy $\eta : \{0, 1\} \times \{0, 1\} \rightarrow \{a, b, c, d\}$ and data encoding policies $g^1 : \{A, B\} \rightarrow \{0, 1\}$ and $g^2 : \{A, B\} \rightarrow \{0, 1\}$ is equivalent to the selection of a mapping $\zeta : \{A, B\} \times \{A, B\} \rightarrow \{a, b, c, d\}$ —because

$$(A.9) \quad \hat{\omega} = \eta(v^1, v^2) = \eta(g^1(y^1), g^2(y^2))$$

—to establish that no sequential design is optimal, it suffices to show that, as long as the mode of at least one of the detectors' sensors is fixed a priori, there is no way that the other detector can program its sensor (in either configuration) so as to ensure that every uncertain outcome $\omega \in \{a, b, c, d\}$ induces a unique element (y^1, y^2)

TABLE A.1
A (graphical) proof that sequential designs are suboptimal.

	configuration 1 mode 0	configuration 1 mode 1	configuration 2 mode 0	configuration 2 mode 1
	 =A  =B	 =A  =B	 =A  =B	 =A  =B
configuration 1 mode 0	a b d	a d	b d	a b
configuration 1 mode 1	a d	a c d	c d	a c
configuration 2 mode 0	b d	c d	b c d	b c
configuration 2 mode 1	a b	a c	b c	a b c

in $\{A, B\} \times \{A, B\}$. Since each sensor has two configurations and two modes, and since there are two detectors, there are 16 cases to consider (eight if we exploit the fact that the sensors available to each detector are identical).

These 16 cases are succinctly summarized in Table A.1. The table can be read as follows. The rows correspond to the possible sensor configurations and fixed modes of the detector that is constrained to act first. The columns correspond to the possible sensor configurations and modes that can be associated with the first detector's uncertain event (i.e., $\{a, b, d\}$ in row 1, $\{a, c, d\}$ in row 2, etc.) when the second detector's sensor configuration and sensor programming policy are appropriately chosen. The table entries correspond to those uncertain outcomes that cannot be distinguished under the stated conditions (i.e., those outcomes that cannot be associated with a unique element of $\{A, B\} \times \{A, B\}$). For example, suppose

- (1) That detector D1 is constrained to use sensor configuration 1 mode 0,
 - (2) That detector D2 uses sensor configuration 2, and
 - (3) That the composition of D2's programming policy with D1's encoding policy (i.e., $f^2 \circ g^1$) maps event B (D1's uncertain event) to mode 0.
- Then, as one can easily verify, the uncertain outcomes b and d are indistinguishable (row one, column three).

Since there is an entry for every possible combination of sensor configurations and modes, under all circumstances, at least two uncertain outcomes are indistinguishable. It follows that no sequential design permits the coordinator to correctly identify, with certainty, more than two of the four uncertain outcomes $\omega \in \{a, b, c, d\}$.

A.4. Summary. By example, it has been shown that nonsequentiality can, under some circumstances, give rise to deadlocks (§A.3), and under other circumstances, improve network performance (§A.4).

Appendix B: Proof of Theorem 1. *Proof of (i).* Fix $\gamma \in \Gamma$ and suppose that ψ is an order function such that \mathcal{I} possesses property CI. To prove that \mathcal{I} possesses property SM it suffices to show that the closed-loop equation $u = \gamma(\omega, u)$ possesses at least one solution, $u_\omega^\gamma \in U$, for all $\omega \in \Omega$; that for each ω , this solution is unique; and that the mapping $\Sigma^\gamma : \Omega \rightarrow U$, induced by these unique solutions (i.e., $\Sigma^\gamma(\omega) = u_\omega^\gamma$) is \mathcal{B}/\mathcal{U} -measurable (see §3.1).

Existence. Fix $\omega \in \Omega$ and $r \in U$. Let π_U denote the canonical projection of $\Omega \times U$ onto U , let $L^\gamma : \Omega \times U \rightarrow \Omega \times U$ be defined as

$$(B.1) \quad L^\gamma(\omega, r) := (\omega, \gamma(\omega, r)),$$

let $L_k^\gamma : \Omega \times U \rightarrow \Omega \times U$ be a k -fold composition of L^γ ,

$$(B.2) \quad L_k^\gamma(\omega, r) := \underbrace{(L^\gamma \circ \dots \circ L^\gamma)}_{[k \text{ times}]}(\omega, r),$$

and let

$$(B.3) \quad s := (s_1, s_2, \dots, s_N) = \psi(L_N^\gamma(\omega, r)).$$

To establish the existence of a closed-loop solution $u_\omega^\gamma \in U$, it suffices to show that

$$(B.4) \quad \begin{aligned} \gamma(L_N^\gamma(\omega, r)) &= \pi_U(L_N^\gamma(\omega, r)) \\ &= \pi_U(\omega, \gamma(L_{N-1}^\gamma(\omega, r))) \\ &= \gamma(L_{N-1}^\gamma(\omega, r)), \end{aligned}$$

or, equivalently, that

$$(B.5) \quad \gamma^{s_k}(L_N^\gamma(\omega, r)) = \gamma^{s_k}(L_{N-1}^\gamma(\omega, r))$$

for all $k = 1, 2, \dots, N$.

Since property CI holds with order function ψ , for all $k = 1, 2, \dots, N$,

$$(B.6) \quad \mathcal{J}^{s_k} \cap [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(L_N^\gamma(\omega, r))) \subset \{\emptyset, [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(L_N^\gamma(\omega, r)))\}.$$

Since \mathcal{U}^k contains the singletons of U^k for all $k = 1, 2, \dots, N$, (B.6) implies that, at the point $L_N^\gamma(\omega, r) \in \Omega \times U$, all $\mathcal{J}^{s_k}/\mathcal{U}^{s_k}$ -measurable functions, including γ^{s_k} , do not depend on components $(s_k + 1), (s_{k+1} + 1), \dots$, and $(s_N + 1)$ of $L_N^\gamma(\omega, r)$; consequently, to establish (B.5) it suffices to show that components 1, $(s_1 + 1), (s_2 + 1), \dots$, and $(s_{k-1} + 1)$ of $L_N^\gamma(\omega, r)$ and $L_{N-1}^\gamma(\omega, r)$ are identical—i.e., it suffices to show that

$$(B.7) \quad \mathcal{P}_{T_{k-1}^N(s)}(L_N^\gamma(\omega, r)) = \mathcal{P}_{T_{k-1}^N(s)}(L_{N-1}^\gamma(\omega, r)).$$

When $k = 1, T_{k-1}^N(s) = \emptyset$, and

$$(B.8) \quad \begin{aligned} \mathcal{P}_\emptyset(L_N^\gamma(\omega, r)) &= \mathcal{P}_\emptyset(\omega, \gamma(L_{N-1}^\gamma(\omega, r))) \\ &= (\omega) \\ &= \mathcal{P}_\emptyset(\omega, \gamma(L_{N-2}^\gamma(\omega, r))) \\ &= \mathcal{P}_\emptyset(L_{N-1}^\gamma(\omega, r)). \end{aligned}$$

For $k > 1$, suppose that (B.7) holds. Then, due to (B.6), (B.5) holds; accordingly,

$$\begin{aligned}
 \mathcal{P}_{T_k^N(s)}(L_N^\gamma(\omega, r)) &= (\mathcal{P}_{T_{k-1}^N(s)}(L_N^\gamma(\omega, r)), \gamma^{s_k}(L_N^\gamma(\omega, r))) \\
 \text{(B.9)} \qquad \qquad \qquad &= (\mathcal{P}_{T_{k-1}^N(s)}(L_{N-1}^\gamma(\omega, r)), \gamma^{s_k}(L_{N-1}^\gamma(\omega, r))) \\
 &= \mathcal{P}_{T_k^N(s)}(L_{N-1}^\gamma(\omega, r)).
 \end{aligned}$$

It follows, by induction, that (B.7) holds for all $k = 1, 2, \dots, N$; hence, (B.5) holds for all $k = 1, 2, \dots, N$, and consequently, (B.4) holds—i.e., $\pi_U(L_N^\gamma(\omega, r))$ satisfies the closed-loop equation.

Uniqueness. Fix $\omega \in \Omega$ and $r \in U$, and once again, let

$$\text{(B.10)} \qquad \qquad \qquad s := (s_1, s_2, \dots, s_N) = \psi(L_N^\gamma(\omega, r)).$$

To establish that $\pi_U(L_N^\gamma(\omega, r))$ is the unique solution to the closed-loop equation $u = \gamma(\omega, u)$ it suffices to show that, $L_N^\gamma(\omega, r) = L_N^\gamma(\omega, \bar{r})$ for all $\bar{r} \in U$, or, equivalently, that

$$\text{(B.11)} \qquad \qquad \qquad \mathcal{P}_{T_{k-1}^N(s)}(L_N^\gamma(\omega, r)) = \mathcal{P}_{T_{k-1}^N(s)}(L_N^\gamma(\omega, \bar{r}))$$

when $k = N + 1$. When $k = 1$, $T_{k-1}^N(s) = \emptyset$, and

$$\begin{aligned}
 \mathcal{P}_\emptyset(L_N^\gamma(\omega, r)) &= \mathcal{P}_\emptyset(\omega, \gamma(L_{N-1}^\gamma(\omega, r))) \\
 \text{(B.12)} \qquad \qquad &= (\omega) \\
 &= \mathcal{P}_\emptyset(\omega, \gamma(L_{N-1}^\gamma(\omega, \bar{r}))) \\
 &= \mathcal{P}_\emptyset(L_N^\gamma(\omega, \bar{r})).
 \end{aligned}$$

For $k > 1$, suppose that (B.11) holds. Then, just as (B.6) and (B.7) imply (B.5), (B.6) and (B.11) imply that

$$\text{(B.13)} \qquad \qquad \qquad \gamma^{s_k}(L_N^\gamma(\omega, r)) = \gamma^{s_k}(L_N^\gamma(\omega, \bar{r}));$$

accordingly,

$$\begin{aligned}
 \mathcal{P}_{T_k^N(s)}(L_N^\gamma(\omega, r)) &= (\mathcal{P}_{T_{k-1}^N(s)}(L_N^\gamma(\omega, r)), \gamma^{s_k}(L_N^\gamma(\omega, r))) \\
 \text{(B.14)} \qquad \qquad \qquad &= (\mathcal{P}_{T_{k-1}^N(s)}(L_N^\gamma(\omega, \bar{r})), \gamma^{s_k}(L_N^\gamma(\omega, \bar{r}))) \\
 &= \mathcal{P}_{T_k^N(s)}(L_N^\gamma(\omega, \bar{r})).
 \end{aligned}$$

It follows, by induction, that (B.11) holds for all $k = 1, 2, \dots, N+1$; hence, $L_N^\gamma(\omega, r) = L_N^\gamma(\omega, \bar{r})$ for all $\bar{r} \in U$, and consequently, the unique solution u_ω^γ to the closed-loop equation $u = \gamma(\omega, u)$ is $\pi_U(L_N^\gamma(\omega, r))$, where $r \in U$ is the (arbitrary) “seed” that starts the recursive solution process.

Measurability. Fix $r \in U$ and let π_U and π_Ω denote, respectively, the canonical projections of $\Omega \times U$ onto U and Ω . To establish the \mathcal{B}/\mathcal{U} -measurability of the induced closed-loop solution map $\Sigma^\gamma : \Omega \rightarrow U$, it suffices to show that the u -section of $\pi_U \circ L_N^\gamma$, $\pi_U \circ L_N^\gamma|_r$, is \mathcal{B}/\mathcal{U} -measurable—because, for fixed r ,

$$\text{(B.15)} \qquad \qquad \qquad \Sigma^\gamma(\omega) = (\pi_U \circ L_N^\gamma|_r)(\omega) := (\pi_U \circ L_N^\gamma)(\omega, r).$$

To begin, note that (B.1) implies that

$$\text{(B.16)} \qquad \qquad \qquad L^\gamma(\omega, r) = (\pi_\Omega(\omega, u), \gamma(\omega, r)).$$

By definition, π_Ω and π_U are, respectively, $\mathcal{B} \otimes \mathcal{U}/\mathcal{B}$ - and $\mathcal{B} \otimes \mathcal{U}/\mathcal{U}$ -measurable. Likewise, $\gamma^k, k = 1, 2, \dots, N$, is $\mathcal{J}^k/\mathcal{U}^k$ -measurable, accordingly, $\gamma := (\gamma^1, \gamma^2, \dots, \gamma^N)$ is $\mathcal{B} \otimes \mathcal{U}/\mathcal{U}$ -measurable (since $\mathcal{J}^k \subset \mathcal{B} \otimes \mathcal{U}$ for all k). It follows that L^γ , and by composition ([7, Thm. 13.1]), L_k^γ ((B.2)) and $\pi_U \circ L_N^\gamma$, are, respectively, $\mathcal{B} \otimes \mathcal{U}/\mathcal{B} \otimes \mathcal{U}$ -, $\mathcal{B} \otimes \mathcal{U}/\mathcal{B} \otimes \mathcal{U}$ -, and $\mathcal{B} \otimes \mathcal{U}/\mathcal{U}$ -measurable. But all u -sections of $\mathcal{B} \otimes \mathcal{U}/\mathcal{U}$ -measurable functions are \mathcal{B}/\mathcal{U} -measurable ([7, Thm. 18.1]); consequently, $\Sigma^\gamma = \pi_U \circ L_N^\gamma|_r$ is \mathcal{B}/\mathcal{U} -measurable.

Proof of (ii).

Sufficiency. Fix $\gamma \in \Gamma$, and suppose that ψ is an order function such that \mathcal{I} possesses property CI. To prove that γ possesses property DF, it suffices to show that for each $\omega \in \Omega$, the agents can be ordered, such that no agent's decision depends on itself or the decisions of its successors.

Fix $\omega \in \Omega$. By (i), the closed-loop equation $u = \gamma(\omega, u)$ possesses a unique solution $u_\omega^\gamma \in U$. Let

$$(B.17) \quad s := (s_1, s_2, \dots, s_N) = \psi(\omega, u_\omega^\gamma).$$

Since property CI holds with order function ψ , for all $k = 1, 2, \dots, N$,

$$(B.18) \quad \mathcal{J}^{s_k} \cap [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega, u_\omega^\gamma)) \subset \{\emptyset, [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega, u_\omega^\gamma))\}.$$

But (B.18) implies that, at the point $(\omega, u_\omega^\gamma) \in \Omega \times U$, all $\mathcal{J}^{s_k}/\mathcal{U}^{s_k}$ -measurable functions, including γ^{s_k} , do not depend on components $(s_k+1), (s_{k+1}+1), \dots$, and (s_N+1) of $(\omega, u_\omega^\gamma)$; consequently, for all $k = 1, 2, \dots, N$, the s_k th agent's decision does not depend on the decisions of agents s_k, s_{k+1}, \dots , and s_N . This proves sufficiency.

Necessity. Suppose that \mathcal{I} does not possess property CI for any order function ψ . Then there exists at least one outcome in $\Omega \times U$, say (ω^*, u^*) , such that for all N -agent orderings $s := (s_1, s_2, \dots, s_N) \in S_N$,

$$(B.19) \quad \mathcal{J}^{s_k} \cap [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega^*, u^*)) \subset \{\emptyset, [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega^*, u^*))\}$$

fails for at least one $k \in \{1, 2, \dots, N\}$. To prove necessity, it suffices to construct a design $\gamma \in \Gamma$ that does not possess property DF.

For all $s \in S_N$, and $k = 1, 2, \dots, N$, let

$$(B.20) \quad \mathcal{L}_s^k := \left\{ A \in \mathcal{J}^{s_k} : (\omega^*, u^*) \in A, A \cap C_s^k(\omega^*, u^*) \notin \{\emptyset, C_s^k(\omega^*, u^*)\} \right\},$$

where

$$(B.21) \quad C_s^k(\omega^*, u^*) := [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega^*, u^*)).$$

When (B.19) holds, $\mathcal{L}_s^k = \emptyset$. When (B.19) fails, \mathcal{L}_s^k contains those events in \mathcal{J}^{s_k} that contain (ω^*, u^*) and depend on the decisions of agents that have yet to act under the decision order s —i.e., those events containing (ω^*, u^*) that, under the decision order s , cannot be distinguished without precognition.

For all $s \in S_N$, and $k = 1, 2, \dots, N$, set

$$(B.22) \quad A_s^{s_k} = U^{s_k} \text{ when } \mathcal{L}_s^k = \emptyset,$$

set

$$(B.23) \quad A_s^{s_k} = A, \quad A \in \mathcal{L}_s^k, \quad A \neq \emptyset, \text{ when } \mathcal{L}_s^k \neq \emptyset,$$

let $r^k \neq u^{*k}$ be an arbitrary reference element in U^k (such an r^k exists since $\text{card}(U^k) > 1$), and let

$$(B.24) \quad \gamma^k(\omega, u) := \begin{cases} u^{*k} & (\omega, u) \in \bigcap_{s' \in S_N} A_{s'}^k, \\ r^k & \text{else.} \end{cases}$$

Since $\text{card}(S_N) = N!$, and $A_s^k \in \mathcal{J}^k$ for all $s \in S_N$, $\bigcap_{s' \in S_N} A_{s'}^k$ is \mathcal{J}^k -measurable; accordingly, γ^k is a $\mathcal{J}^k/\mathcal{U}^k$ -measurable function for all $k = 1, 2, \dots, N$, and consequently, $\gamma := (\gamma^1, \gamma^2, \dots, \gamma^N)$ is an element of Γ .

The design γ , however, is not deadlock-free. Consider the outcome (ω^*, u^*) , fix $s \in S_N$, and let k^* denote a k for which (B.19) fails. By construction (ω^*, u^*) satisfies the closed-loop equation (i.e., $u^* = \gamma(\omega^*, u^*)$); moreover, $\mathcal{L}_s^{k^*} \neq \emptyset$. It follows from (B.23) that $A_{s^{k^*}}^s \in \mathcal{L}_s^{k^*}$, and $A_{s^{k^*}}^s \neq \emptyset$; accordingly,

$$(B.25) \quad \begin{aligned} [\gamma^{s_{k^*}}]^{-1}(u^{*s_{k^*}}) \cap C_s^{k^*}(\omega^*, u^*) &= \left(\bigcap_{s' \in S_N} A_{s'}^{s_{k^*}} \right) \cap C_s^{k^*}(\omega^*, u^*) \\ &\not\subseteq \{\emptyset, C_s^{k^*}(\omega^*, u^*)\}. \end{aligned}$$

However, (B.25) implies that, at the point $(\omega^*, u^*) \in \Omega \times U$, agent s_{k^*} 's decision depends on the decision of agents that have yet to act under s . Since the same argument applies for all $s \in S_N$, γ does not possess property DF. This proves necessity. \square

Appendix C: Proof of Corollary 1. Although this corollary is an immediate consequence of Theorems 2(ii) and 1(ii) (property C \Rightarrow property DF \Rightarrow property CI), it is instructive to prove it directly.

Suppose that ψ is an order function such that \mathcal{I} possesses property C. It suffices to show that ψ is also an order function such that \mathcal{I} possesses property CI—i.e., that (4.1) of property C (with $s = T_k^N(\psi(\omega, u)) \in S_k$), implies (3.3) of property CI (with $s = \psi(\omega, u) \in S_N$), for all $(\omega, u) \in \Omega \times U$ and $k = 1, 2, \dots, N$.

Fix $(\omega, u) \in \Omega \times U$ and $k \in \{1, 2, \dots, N\}$, and let

$$(C.1) \quad s := (s_1, s_2, \dots, s_N) = \psi(\omega, u).$$

Since $T_k^N(s) \in S_k$, and $T_{k-1}^N = T_{k-1}^k \circ T_k^N$, (4.1) of property C implies that

$$(C.2) \quad \mathcal{J}^{s_k} \cap [T_k^N \circ \psi]^{-1}(T_k^N(s)) \subset \mathcal{F}(T_{k-1}^N(s)).$$

Restricting both sides of (C.2) to

$$(C.3) \quad [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega, u))$$

yields the desired result—(3.3) of property CI—if

$$(C.4) \quad \begin{aligned} [T_k^N \circ \psi]^{-1}(T_k^N(s)) \cap [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega, u)) \\ = [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega, u)) \end{aligned}$$

and

$$(C.5) \quad \begin{aligned} \mathcal{F}(T_{k-1}^N(s)) \cap [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega, u)) \\ = \{\emptyset, [\mathcal{P}_{T_{k-1}^N(s)}]^{-1}(\mathcal{P}_{T_{k-1}^N(s)}(\omega, u))\}. \end{aligned}$$

Equation (C.5) follows from the definition of $\mathcal{F}(T_{k-1}^N(s))$,

$$(C.6) \quad \mathcal{F}(T_{k-1}^N(s)) := [\mathcal{P}_{T_{k-1}^N(s)}]^{-1} \left(\mathcal{B} \otimes \left(\bigotimes_{i=1}^{k-1} \mathcal{U}^{s_i} \right) \right),$$

and the fact that inverse images preserve intersections—i.e.,

$$(C.7) \quad [\mathcal{P}_{T_{k-1}^N(s)}]^{-1} \left(\mathcal{B} \otimes \left(\bigotimes_{i=1}^{k-1} \mathcal{U}^{s_i} \right) \right) \cap [\mathcal{P}_{T_{k-1}^N(s)}]^{-1} (\mathcal{P}_{T_{k-1}^N(s)}(\omega, u)) \\ = \{\emptyset, [\mathcal{P}_{T_{k-1}^N(s)}]^{-1} (\mathcal{P}_{T_{k-1}^N(s)}(\omega, u))\}.$$

Equation (C.4) follows from the observation that

$$(C.8) \quad [T_k^N \circ \psi]^{-1}(T_k^N(s)) \in \mathcal{F}(T_{k-1}^N(s))$$

(to see this substitute $\Omega \times U \in \mathcal{J}^{s_k}$ for \mathcal{J}^{s_k} , and \in for \subset , in (C.2)), (C.5), and the fact that

$$(C.9) \quad [T_k^N \circ \psi]^{-1}(T_k^N(s)) \cap [\mathcal{P}_{T_{k-1}^N(s)}]^{-1} (\mathcal{P}_{T_{k-1}^N(s)}(\omega, u)) \neq \emptyset$$

since both sets contain (ω, u) . \square

Appendix D: Proof of Theorem 3. Suppose that \mathcal{I} is sequential. Then there exists a *constant* order function ψ such that \mathcal{I} possesses property CI. It suffices to show that ψ is also an order function such that \mathcal{I} possesses property C—i.e., that for all $k = 1, 2, \dots, N$, the fact that (3.3) of property CI holds for all $(\omega, u) \in \Omega \times U$ with $s = s^* \in S_N$ constant, implies that (4.1) of property C holds for all $s \in S_k$.

Fix $k \in \{1, 2, \dots, N\}$ and let

$$(D.1) \quad s^* := (s_1^*, s_2^*, \dots, s_N^*)$$

denote the constant order induced by ψ . Since

$$(D.2) \quad [T_k^N \circ \psi]^{-1}(s) = \begin{cases} \Omega \times U & \text{when } s = T_k^N(s^*) \\ \emptyset & \text{else} \end{cases}$$

for all $s \in S_k$, and since $T_{k-1}^N = T_{k-1}^k \circ T_k^N$, to prove that (4.1) of property C holds for all $s \in S_k$, it suffices to show that

$$(D.3) \quad \mathcal{J}^{s_k^*} \subset \mathcal{F}(T_{k-1}^N(s^*)).$$

By definition, $\mathcal{J}^{s_k^*}$ is a subfield of

$$(D.4) \quad \mathcal{B} \otimes \mathcal{U} = [\mathcal{P}_{T_N^N(s^*)}]^{-1} \left(\mathcal{B} \otimes \left(\bigotimes_{i=1}^N \mathcal{U}^{s_i^*} \right) \right).$$

Since (3.3) holds for all $(\omega, u) \in \Omega \times U$ when $s = s^*$, all events in $\mathcal{J}^{s_k^*}$ must be of the form

$$(D.5) \quad [\mathcal{P}_{T_N^N(s^*)}]^{-1} \left(A \times \left(\prod_{i=k}^N \mathcal{U}^{s_i^*} \right) \right),$$

where $A \subset \Omega \times \prod_{i=1}^{k-1} U^{s_i^*}$; accordingly, \mathcal{J}^{s^*} is also a subfield of

$$\begin{aligned} \mathcal{C}_{s^*} &:= \sigma \left([\mathcal{P}_{T_N^N(s^*)}]^{-1} \left(A \times \left(\prod_{i=k}^N U^{s_i^*} \right) \right) : A \subset \Omega \times \prod_{i=1}^{k-1} U^{s_i^*} \right) \\ (D.6) \quad &= \sigma \left([\mathcal{P}_{T_{k-1}^N(s^*)}]^{-1}(A) : A \subset \Omega \times \prod_{i=1}^{k-1} U^{s_i^*} \right) \end{aligned}$$

—the cylindrical extension of the power set of $\Omega \times \prod_{i=1}^{k-1} U^{s_i^*}$ to $\Omega \times U$. However,

$$\begin{aligned} (\mathcal{B} \otimes \mathcal{U}) \cap \mathcal{C}_{s^*} &= [\mathcal{P}_{T_{k-1}^N(s^*)}]^{-1} \left(\mathcal{B} \otimes \left(\bigotimes_{i=1}^{k-1} \mathcal{U}^{s_i^*} \right) \right) \\ (D.7) \quad &:= \mathcal{F}(T_{k-1}^N(s^*)); \end{aligned}$$

consequently, $\mathcal{J}^{s^*} \subset \mathcal{F}(T_{k-1}^N(s^*))$. \square

Appendix E: Proof of Theorem 4. Suppose that ψ is an order function such that \mathcal{I} possesses property CI. It suffices to construct an order function $\hat{\psi}$ such that \mathcal{I} possesses property C.

To simplify property C's verification, it is convenient to construct $\hat{\psi}$ recursively. The recursion has N steps, the k th of which, $k = 1, 2, \dots, N$, corresponds to the construction of a function

$$(E.1) \quad f_k : \Omega \times U \rightarrow S_k$$

with the following properties:

- (1) For all $j \in \{1, 2, \dots, k-1\}$, $T_j^k \circ f_k = f_j$, and
- (2) For all $s := (s_1, s_2, \dots, s_k) \in S_k$, $\mathcal{J}^{s_k} \cap [f_k]^{-1}(s) \subset \mathcal{F}(T_{k-1}^k(s))$.

Property (1) suffices to ensure that $f_k = [T_k^N \circ f_N]$; consequently, property (2) suffices to ensure that $\hat{\psi} = f_N$ is an order function such that \mathcal{I} possesses property C (see Definition 3 in §4).

For all $(\omega, u) \in \Omega \times U$,

$$(E.2) \quad s := (s_1, s_2, \dots, s_{k-1}) \in S_{k-1},$$

and $k = 1, 2, \dots, N$: let

$$(E.3) \quad C_s(\omega, u) := [\mathcal{P}_s]^{-1}(\mathcal{P}_s(\omega, u))$$

denote the cylinder set induced on $\Omega \times U$ by $(\omega, u^{s_1}, \dots, u^{s_{k-1}})$; let

$$(E.4) \quad \langle\langle s \rangle\rangle := (\langle\langle s \rangle\rangle_1, \langle\langle s \rangle\rangle_2, \dots, \langle\langle s \rangle\rangle_N) \in S_N$$

denote the unique element in S_N for which $T_{k-1}^N(\langle\langle s \rangle\rangle) = s$, and $\langle\langle s \rangle\rangle_k < \langle\langle s \rangle\rangle_{k+1} < \dots < \langle\langle s \rangle\rangle_N$; and let

$$(E.5) \quad s, \langle\langle s \rangle\rangle_j := (s_1, s_2, \dots, s_{k-1}, \langle\langle s \rangle\rangle_j)$$

$j = k, k+1, \dots, N$, denote the concatenation of $\langle\langle s \rangle\rangle_j$ to s . Then the recursive construction of $\hat{\psi}$, given ψ , can be described as follows:

1. For all $j = 1, 2, \dots, N$, let

$$(E.6) \quad f_1(\omega, u) = j$$

when

$$(E.7) \quad (\omega, u) \in C_\emptyset([T_1^N \circ \psi]^{-1}(j)) \setminus \left(\bigcup_{i=1}^{j-1} C_\emptyset([T_1^N \circ \psi]^{-1}(i)) \right).^{10}$$

\vdots

k. For all $s \in S_{k-1}$, and $j = k, k+1, \dots, N$, let

$$(E.8) \quad f_k(\omega, u) = s, \langle \langle s \rangle \rangle_j$$

when

$$(E.9) \quad (\omega, u) \in [f_{k-1}]^{-1}(s) \cap \left(C_s([T_k^N \circ \psi]^{-1}(s, \langle \langle s \rangle \rangle_j)) \setminus \left(\bigcup_{i=k}^{j-1} C_s([T_k^N \circ \psi]^{-1}(s, \langle \langle s \rangle \rangle_i)) \right) \right).$$

\vdots

N. For all $s \in S_{N-1}$, let

$$(E.10) \quad f_N(\omega, u) = s, \langle \langle s \rangle \rangle_N$$

when

$$(E.11) \quad (\omega, u) \in [f_{N-1}]^{-1}(s) \cap C_s([T_{N-1}^N \circ \psi]^{-1}(s, \langle \langle s \rangle \rangle_N)).$$

To verify that the preceding constructions give rise to legitimate functions it suffices to check, for all $k = 1, 2, \dots, N$, that $\{[f_k]^{-1}(s) : s \in S_k\}$ partitions $\Omega \times U$. The following facts will be used without comment:

- Unions and intersections are distributive.
- Inverse and direct images preserve unions and inclusions.
- $\{[T_k^N \circ \psi]^{-1}(s) : s \in S_k\}$ partitions $\Omega \times U$ for all $k = 1, 2, \dots, N$; moreover, since

$$(E.12) \quad [T_{k-1}^k]^{-1}(s) = \bigcup_{i=k}^N (s, \langle \langle s \rangle \rangle_i),$$

for all $s \in S_{k-1}$, $k = 1, 2, \dots, N$,

$$(E.13) \quad \begin{aligned} [T_{k-1}^N \circ \psi]^{-1}(s) &= [T_{k-1}^k \circ T_k^N \circ \psi]^{-1}(s) \\ &= [T_k^N \circ \psi]^{-1}([T_{k-1}^k]^{-1}(s)) \\ &= [T_k^N \circ \psi]^{-1} \left(\bigcup_{i=k}^N (s, \langle \langle s \rangle \rangle_i) \right) \\ &= \bigcup_{i=k}^N [T_k^N \circ \psi]^{-1}(s, \langle \langle s \rangle \rangle_i). \end{aligned}$$

¹⁰ For sets $A, B \in X$, $A \setminus B := \{x \in A : x \notin B\}$.

- When A, B, C, D, E are sets,

$$A \cup (B \setminus A) = A \cup B, \quad A \cup B \cup (C \setminus (A \cup B)) = A \cup B \cup C, \text{ and so on, and}$$

$$(E.14) \quad A \cap (B \setminus A) = \emptyset, \quad C \cap (E \setminus (A \cup B \cup C \cup D)) = \emptyset, \text{ and so on.}$$

When $k = 1$,

$$\begin{aligned}
 [f_1]^{-1}(S_1) &:= [f_1]^{-1} \left(\bigcup_{j=1}^N \{j\} \right) \\
 &= \bigcup_{j=1}^N [f_1]^{-1}(j) \\
 &:= \bigcup_{j=1}^N \left(C_{\emptyset}([T_1^N \circ \psi]^{-1}(j)) \setminus \left(\bigcup_{i=1}^{j-1} C_{\emptyset}([T_1^N \circ \psi]^{-1}(i)) \right) \right) \\
 (E.15) \quad &= \bigcup_{j=1}^N C_{\emptyset}([T_1^N \circ \psi]^{-1}(j)) \\
 &= C_{\emptyset} \left(\bigcup_{j=1}^N [T_1^N \circ \psi]^{-1}(j) \right) \\
 &= C_{\emptyset}(\Omega \times U) \\
 &\supset \Omega \times U.
 \end{aligned}$$

Moreover, (E.6) and (E.7) imply that for all $m, n \in \{1, 2, \dots, N\}, m < n$,

$$\begin{aligned}
 [f_1]^{-1}(m) \cap [f_1]^{-1}(n) &:= \left(C_{\emptyset}([T_1^N \circ \psi]^{-1}(m)) \setminus \left(\bigcup_{i=1}^{m-1} C_{\emptyset}([T_1^N \circ \psi]^{-1}(i)) \right) \right) \\
 &\quad \cap \left(C_{\emptyset}([T_1^N \circ \psi]^{-1}(n)) \setminus \left(\bigcup_{i=1}^{n-1} C_{\emptyset}([T_1^N \circ \psi]^{-1}(i)) \right) \right) \\
 (E.16) \quad &\subset C_{\emptyset}([T_1^N \circ \psi]^{-1}(m)) \\
 &\quad \cap \left(C_{\emptyset}([T_1^N \circ \psi]^{-1}(n)) \setminus \left(\bigcup_{i=1}^{n-1} C_{\emptyset}([T_1^N \circ \psi]^{-1}(i)) \right) \right) \\
 &= \emptyset.
 \end{aligned}$$

It follows that $\{[f_1]^{-1}(s) : s \in S_1\}$ partitions $\Omega \times U$.

For $k > 1$, suppose that $\{[f_{k-1}]^{-1}(s) : s \in S_{k-1}\}$ partitions $\Omega \times U$. Then

$$\begin{aligned}
 [f_k]^{-1}(S_k) &:= [f_k]^{-1} \left(\bigcup_{s' \in S_k} s' \right) \\
 &:= [f_k]^{-1} \left(\bigcup_{s \in S_{k-1}} \bigcup_{j=k}^N (s, \langle \langle s \rangle \rangle_j) \right) \\
 &= \bigcup_{s \in S_{k-1}} \bigcup_{j=k}^N [f_k]^{-1}(s, \langle \langle s \rangle \rangle_j)
 \end{aligned}$$

$$\begin{aligned}
&:= \bigcup_{s \in S_{k-1}} \bigcup_{j=k}^N \left([f_{k-1}]^{-1}(s) \cap \left(C_s([T_k^N \circ \psi]^{-1}(s, \langle \langle s \rangle \rangle_j)) \right. \right. \\
&\quad \left. \left. \setminus \left(\bigcup_{i=k}^{j-1} C_s([T_k^N \circ \psi]^{-1}(s, \langle \langle s \rangle \rangle_i)) \right) \right) \right) \\
&= \bigcup_{s \in S_{k-1}} \left([f_{k-1}]^{-1}(s) \cap \left(\bigcup_{j=k}^N \left(C_s([T_k^N \circ \psi]^{-1}(s, \langle \langle s \rangle \rangle_j)) \right. \right. \right. \\
&\quad \left. \left. \setminus \left(\bigcup_{i=k}^{j-1} C_s([T_k^N \circ \psi]^{-1}(s, \langle \langle s \rangle \rangle_i)) \right) \right) \right) \quad (\text{E.17}) \\
&= \bigcup_{s \in S_{k-1}} \left([f_{k-1}]^{-1}(s) \cap \left(\bigcup_{j=k}^N C_s([T_k^N \circ \psi]^{-1}(s, \langle \langle s \rangle \rangle_j)) \right) \right) \\
&= \bigcup_{s \in S_{k-1}} \left([f_{k-1}]^{-1}(s) \cap C_s \left(\bigcup_{j=k}^N [T_k^N \circ \psi]^{-1}(s, \langle \langle s \rangle \rangle_j) \right) \right) \\
&= \bigcup_{s \in S_{k-1}} \left([f_{k-1}]^{-1}(s) \cap C_s([T_{k-1}^N \circ \psi]^{-1}(s)) \right) \\
&= \left(\bigcup_{s \in S_{k-1}} [f_{k-1}]^{-1}(s) \right) \cap \left(\bigcup_{s \in S_{k-1}} C_s([T_{k-1}^N \circ \psi]^{-1}(s)) \right) \\
&= (\Omega \times U) \cap \left(\bigcup_{s \in S_{k-1}} C_s([T_{k-1}^N \circ \psi]^{-1}(s)) \right) \\
&= \bigcup_{s \in S_{k-1}} C_s([T_{k-1}^N \circ \psi]^{-1}(s)) \\
&\supset \bigcup_{s \in S_{k-1}} [T_{k-1}^N \circ \psi]^{-1}(s) \\
&= \Omega \times U.
\end{aligned}$$

Moreover, for all $s, \bar{s} \in S_k$ such that $s \neq \bar{s}$, when $T_{k-1}^k(s) \neq T_{k-1}^k(\bar{s})$, (E.8) and (E.9) and the induction hypothesis imply that

$$\begin{aligned}
[f_k]^{-1}(s) \cap [f_k]^{-1}(\bar{s}) &\subset [f_{k-1}]^{-1}(T_{k-1}^k(s)) \cap [f_{k-1}]^{-1}(T_{k-1}^k(\bar{s})) \\
&= \emptyset, \quad (\text{E.18})
\end{aligned}$$

and when $T_{k-1}^k(s) = T_{k-1}^k(\bar{s})$ (implying that $s_k \neq \bar{s}_k$), (E.8) and (E.9) and the induction hypothesis imply that for some $m < n$ (say $s_k = \langle \langle s \rangle \rangle_m, \bar{s}_k = \langle \langle \bar{s} \rangle \rangle_n$)

$$\begin{aligned}
[f_k]^{-1}(s) \cap [f_k]^{-1}(\bar{s}) &\subset \left(C_s([T_k^N \circ \psi]^{-1}(s, \langle \langle s \rangle \rangle_m)) \right. \\
&\quad \left. \setminus \left(\bigcup_{i=k}^{m-1} C_s([T_k^N \circ \psi]^{-1}(s, \langle \langle s \rangle \rangle_i)) \right) \right) \\
&\quad \cap \left(C_s([T_k^N \circ \psi]^{-1}(s, \langle \langle s \rangle \rangle_n)) \right)
\end{aligned}$$

$$\begin{aligned}
& \setminus \left(\bigcup_{i=k}^{n-1} C_s([T_k^N \circ \psi]^{-1}(s, \langle\langle s \rangle\rangle_i)) \right) \\
& \subset C_s([T_k^N \circ \psi]^{-1}(s, \langle\langle s \rangle\rangle_m)) \\
(E.19) \quad & \cap \left(C_s([T_k^N \circ \psi]^{-1}(s, \langle\langle s \rangle\rangle_n)) \right. \\
& \left. \setminus \left(\bigcup_{i=k}^{n-1} C_s([T_k^N \circ \psi]^{-1}(s, \langle\langle s \rangle\rangle_i)) \right) \right) \\
& = \emptyset.
\end{aligned}$$

Consequently, $\{[f_k]^{-1}(s) : s \in S_k\}$ partitions $\Omega \times U$. It follows, by induction, that for all $k = 1, 2, \dots, N$, $\{[f_k]^{-1}(s) : s \in S_k\}$ partitions $\Omega \times U$.

Having established, for all $k = 1, 2, \dots, N$, that f_k is a legitimate function, it remains to show that f_k satisfies properties (1) and (2) (cf. the discussion following (E.1)). To verify property (1) it suffices to prove, for all $k = 1, 2, \dots, N$, that

$$(E.20) \quad T_{k-1}^k \circ f_k = f_{k-1},$$

or, equivalently, that

$$(E.21) \quad [T_{k-1}^k \circ f_k]^{-1}(s) = [f_{k-1}]^{-1}(s)$$

for all $s \in S_{k-1}$. Fix $k \in \{1, 2, \dots, N\}$. By (E.8) and (E.9)

$$(E.22) \quad [f_k]^{-1}(s, \langle\langle s \rangle\rangle_j) \subset [f_{k-1}]^{-1}(s)$$

for all $s \in S_{k-1}$ and $j = k, k+1, \dots, N$; consequently, since $\{[f_{k-1}]^{-1}(s) : s \in S_{k-1}\}$ partitions $\Omega \times U$, for all $s, \bar{s} \in S_{k-1}$ such that $s \neq \bar{s}$, and for arbitrary $j \in \{k, k+1, \dots, N\}$,

$$(E.23) \quad [f_k]^{-1}(s, \langle\langle s \rangle\rangle_j) \cap [f_{k-1}]^{-1}(\bar{s}) = \emptyset.$$

However, $\{[f_k]^{-1}(s) : s \in S_k\}$ also partitions $\Omega \times U$; accordingly, (E.23) implies that for all $s \in S_{k-1}$,

$$\begin{aligned}
(E.24) \quad [T_{k-1}^k \circ f_k]^{-1}(s) &= [f_k]^{-1} \circ [T_{k-1}^k]^{-1}(s) \\
&= [f_k]^{-1} \left(\bigcup_{i=k}^N (s, \langle\langle s \rangle\rangle_i) \right) \\
&= \bigcup_{i=k}^N [f_k]^{-1}(s, \langle\langle s \rangle\rangle_i) \\
&= \left(\bigcup_{\bar{s} \in S_{k-1}} \bigcup_{i=k}^N [f_k]^{-1}(\bar{s}, \langle\langle \bar{s} \rangle\rangle_i) \right) \cap [f_{k-1}]^{-1}(s) \\
&= \left(\bigcup_{s' \in S_k} [f_k]^{-1}(s') \right) \cap [f_{k-1}]^{-1}(s) \\
&= (\Omega \times U) \cap [f_{k-1}]^{-1}(s) \\
&= [f_{k-1}]^{-1}(s)
\end{aligned}$$

—i.e., (E.21), and, consequently, property (1) hold.

To verify property (2) (see the discussion following (E.1)) it is necessary to establish the following lemma.

LEMMA E1. *Suppose that Ω , and $U^k, k = 1, 2, \dots, N$, are countable sets, and suppose that \mathcal{B} contains the singletons of Ω . Then if ψ is an order function such that \mathcal{I} possesses property CI, for all $s \in S_k, k = 1, 2, \dots, N$,*

$$(E.25) \quad \mathcal{J}^{s_k} \cap C_{T_{k-1}^k(s)}([T_k^N \circ \psi]^{-1}(s)) \subset \mathcal{F}(T_{k-1}^k(s)).$$

Proof. By assumption, the σ -fields \mathcal{B} and $\mathcal{U}^k, k = 1, 2, \dots, N$, contain, respectively, the singletons of the countable sets Ω and $U^k, k = 1, 2, \dots, N$ (\mathcal{U}^k contains the singletons of U^k due to §2.2, 1(c)). Accordingly, for all $s := (s_1, s_2, \dots, s_k) \in S_k, k = 1, 2, \dots, N$, the product field $\mathcal{B} \otimes (\bigotimes_{i=1}^k \mathcal{U}^{s_i})$ contains the singletons of the countable set $\Omega \times (\prod_{i=1}^k U^{s_i})$, implying that $\mathcal{B} \otimes (\bigotimes_{i=1}^k \mathcal{U}^{s_i})$ is the power set of $\Omega \times (\prod_{i=1}^k U^{s_i})$. It follows, for all $s \in S_k, k = 1, 2, \dots, N$, that

$$(E.26) \quad \begin{aligned} \mathcal{F}(T_{k-1}^k(s)) &:= [\mathcal{P}_{T_{k-1}^k(s)}]^{-1} \left(\mathcal{B} \otimes \left(\bigotimes_{i=1}^{k-1} \mathcal{U}^{s_i} \right) \right) \\ &= \sigma \left([\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(A) : A \subset \Omega \times \prod_{i=1}^{k-1} U^{s_i} \right) \end{aligned}$$

—i.e., it follows that $\mathcal{F}(T_{k-1}^k(s))$ is the cylindrical extension of the power set of $\Omega \times \prod_{i=1}^{k-1} U^{s_i}$ to $\Omega \times U$.

Fix $k \in \{1, 2, \dots, N\}$ and $s \in S_k$. Since property CI holds with order function ψ , (E.3), (3.3), and (E.26) imply that for all $(\omega, u) \in [T_k^N \circ \psi]^{-1}(s)$ and $A \in \mathcal{J}^{s_k}$,

$$(E.27) \quad \begin{aligned} A \cap C_{T_{k-1}^k(s)}(\omega, u) &= A \cap [\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}(\omega, u)) \\ &\in \{\emptyset, [\mathcal{P}_{T_{k-1}^k(s)}]^{-1}(\mathcal{P}_{T_{k-1}^k(s)}(\omega, u))\} \\ &\subset \mathcal{F}(T_{k-1}^k(s)). \end{aligned}$$

Since $[T_k^N \circ \psi]^{-1}(s) \in \Omega \times U$ is a countable set, and since inverse and direct images preserve unions, it follows that

$$(E.28) \quad \begin{aligned} A \cap C_{T_{k-1}^k(s)}([T_k^N \circ \psi]^{-1}(s)) &= A \cap C_{T_{k-1}^k(s)} \left(\bigcup_{(\omega, u) \in [T_k^N \circ \psi]^{-1}(s)} (\omega, u) \right) \\ &= \bigcup_{(\omega, u) \in [T_k^N \circ \psi]^{-1}(s)} (A \cap C_{T_{k-1}^k(s)}(\omega, u)) \\ &\in \mathcal{F}(T_{k-1}^k(s)). \end{aligned}$$

This proves the lemma since (E.28) holds for all $A \in \mathcal{J}^{s_k}$, and consequently, implies (E.25). \square

Given Lemma E1, by induction, all f_k can be shown to possess property (2). For $k = 1$, fix $j \in \{1, 2, \dots, N\}$. By Lemma E1, for all $A \in \mathcal{J}^j$,

$$(E.29) \quad A \cap C_{\emptyset}([T_1^N \circ \psi]^{-1}(j)) \in \mathcal{F}(\emptyset).$$

Likewise, since $\Omega \times U \in \mathcal{J}^i$ for all i , for all $i = 1, 2, \dots, N$,

$$(E.30) \quad C_{\emptyset}([T_1^N \circ \psi]^{-1}(i)) \in \mathcal{F}(\emptyset);$$

accordingly,

$$(E.31) \quad \bigcup_{i=1}^{j-1} C_{\emptyset}([T_1^N \circ \psi]^{-1}(i)) \in \mathcal{F}(\emptyset).$$

It follows, from (E.29) and (E.31), that

$$(E.32) \quad A \cap [f_1]^{-1}(j) := A \cap \left(C_{\emptyset}([T_1^N \circ \psi]^{-1}(j)) \right. \\ \left. \setminus \left(\bigcup_{i=1}^{j-1} C_{\emptyset}([T_1^N \circ \psi]^{-1}(i)) \right) \right) \\ (E.33) \quad \in \mathcal{F}(\emptyset).$$

Since (E.32) holds for all $A \in \mathcal{J}^j$, f_1 satisfies property (2)—i.e., for all $j \in S_1$,

$$(E.34) \quad \mathcal{J}^j \cap [f_1]^{-1}(j) \subset \mathcal{F}(\emptyset).$$

For $k > 1$, suppose that f_{k-1} satisfies property (2)—i.e., suppose that, for all $s \in S_{k-1}$,

$$(E.35) \quad \mathcal{J}^{s_{k-1}} \cap [f_{k-1}]^{-1}(s) \subset \mathcal{F}(T_{k-2}^{k-1}(s)).$$

Then, since $\Omega \times U \in \mathcal{J}^i$ for all $i = 1, 2, \dots, N$, for all $s \in S_{k-1}$,

$$(E.36) \quad [f_{k-1}]^{-1}(s) \subset \mathcal{F}(T_{k-2}^{k-1}(s)) \subset \mathcal{F}(s).$$

Fix $s \in S_{k-1}$ and $j \in \{k, k+1, \dots, N\}$. By Lemma E1, for all $A \in \mathcal{J}^{\langle\langle s \rangle\rangle_j}$,

$$(E.37) \quad A \cap (C_s([T_k^N \circ \psi]^{-1}(s, \langle\langle s \rangle\rangle_j))) \in \mathcal{F}(s).$$

Likewise, since $\Omega \times U \in \mathcal{J}^i$ for all i , for all $i = k, k+1, \dots, N$,

$$(E.38) \quad C_s([T_k^N \circ \psi]^{-1}(s, \langle\langle s \rangle\rangle_i)) \in \mathcal{F}(s);$$

accordingly,

$$(E.39) \quad \bigcup_{i=k}^{j-1} C_s([T_k^N \circ \psi]^{-1}(s, \langle\langle s \rangle\rangle_i)) \in \mathcal{F}(s).$$

It follows, from (E.35), (E.36), and (E.38), that

$$(E.40) \quad A \cap [f_k]^{-1}(s, \langle\langle s \rangle\rangle_j) := A \cap [f_{k-1}]^{-1}(s) \cap \left(C_s([T_k^N \circ \psi]^{-1}(s, \langle\langle s \rangle\rangle_j)) \right. \\ \left. \setminus \left(\bigcup_{i=k}^{j-1} C_s([T_k^N \circ \psi]^{-1}(s, \langle\langle s \rangle\rangle_i)) \right) \right) \\ \in \mathcal{F}(s).$$

Since (E.39) holds for all $A \in \mathcal{J}^{\langle\langle s \rangle\rangle_j}$, f_k satisfies property (2)—i.e., for all $s \in S_k$

$$(E.41) \quad \mathcal{J}^{s_k} \cap [f_k]^{-1}(s) \subset \mathcal{F}(T_{k-1}^k(s)).$$

It follows, by induction, that f_k satisfies property (2) for all $k = 1, 2, \dots, N$; consequently, since all f_k 's also satisfy property (1), $\hat{\psi} = f_N$ is an order function such that \mathcal{I} possesses property C (see the discussion following (E.1)). This proves the theorem. \square

Acknowledgments. The authors thank the referees and editors for their helpful comments.

REFERENCES

- [1] M. S. ANDERSLAND, *Decoupling nonsequential stochastic control problems*, Systems Control Lett., 16 (1991), pp. 65–69.
- [2] M. S. ANDERSLAND AND D. TENEKETZIS, *Solvable systems are usually measurable*, Stochastic Analysis and Applications, 9 (1991), pp. 233–244.
- [3] ———, *Information structures, causality, and nonsequential stochastic control II: design-dependent properties*, SIAM J. Control Optim., submitted.
- [4] A. BENVENISTE AND P. LE GUERNIC, *Hybrid dynamical system theory and the SIGNAL language*, IEEE Trans. Automat. Control, AC-35 (1990), pp. 533–546.
- [5] P. A. BERNSTEIN, V. HADZILACOS, AND N. GOODMAN, *Concurrency Control and Recovery in Database Systems*, Addison-Wesley, Reading, MA, 1987.
- [6] E. BEST AND C. FERNÁNDEZ, *Nonsequential Processes: A Petri Net View*, Springer-Verlag, Berlin, 1989.
- [7] P. BILLINGSLEY, *Probability and Measure*, 2nd ed., John Wiley, New York, 1986.
- [8] J.Y. HALPERN AND R. FAGIN, *Modeling knowledge and action in distributed systems*, Lecture Notes in Comp. Sci., Vol. 335, Springer-Verlag, Berlin, 1989, pp. 18–32.
- [9] C.A.R. HOARE, *Communicating Sequential Processes*, Prentice Hall, Englewood Cliffs, NJ, 1985.
- [10] K. INAN AND P. VARAIYA, *Finitely recursive process models for discrete event systems*, IEEE Trans. Automat. Control, AC-33, (1988), pp. 626–639.
- [11] Z. BANASZAK AND B.H. KROUGH, *Deadlock avoidance in flexible manufacturing systems with concurrently competing process flows*, IEEE Trans. Robotics Automation, RA-6 (1990), pp. 724–734.
- [12] H.W. KUHN, *Extensive games and the problem of information*, Contributions to the Theory of Games, Vol. 2, Annals of Math. Studies, 28 (1953), pp. 193–216.
- [13] W.S. LAI, *Protocol traps in computer networks—a catalog*, IEEE Trans. Commun., COM-30 (1982), pp. 1434–1449.
- [14] Z. MANNA AND A. PNUELI, *The anchored version of the temporal framework*, Lecture Notes in Comp. Sci., Vol. 354, Springer-Verlag, Berlin, 1989, 201–284.
- [15] R. MILNER, *Communication and Concurrency*, Prentice-Hall, Englewood Cliffs, NJ, 1989.
- [16] P.J. RAMADGE AND W.M. WONHAM, *The control of discrete event systems*, Proc. of the IEEE, 77 (1989), pp. 81–98.
- [17] J.G. THISTLE AND W.M. WONHAM, *Control problems in a temporal logic framework*, Int. J. Control, 44 (1986), pp. 943–976.
- [18] J. VON NEUMANN AND O. MORGENSTERN, *The Theory of Games and Economic Behavior*, Princeton, NJ: Princeton U. Press, Chap. 2, 1944.
- [19] H.S. WITSENHAUSEN, *On information structures, feedback and causality*, SIAM J. Control, 9 (1971), pp. 149–160.
- [20] ———, *A standard form for sequential stochastic control*, Math. Sys. Theory, 7 (1973), pp. 5–11.
- [21] ———, *The intrinsic model for discrete stochastic control: some open problems*, Lecture Notes in Econ. and Math. Sys., Vol. 107, Springer-Verlag, Berlin, 1975, pp. 322–335.

STOCHASTIC APPROXIMATIONS AND ADAPTIVE CONTROL OF A DISCRETE-TIME SINGLE-SERVER NETWORK WITH RANDOM ROUTING*

ARMAND M. MAKOWSKI† AND ADAM SHWARTZ‡

This paper is dedicated to the memory of Michel Metivier.

Abstract. This paper considers a discrete-time system composed of K infinite capacity queues that compete for the use of a single server. Customers arrive in independent and identically distributed (i.i.d.) batches and are served according to a server allocation policy. Upon completing service, customers either leave the system or are routed instantaneously to another queue according to some random mechanism. As an alternative to simply randomized strategies, a policy based on a stochastic approximation algorithm is proposed to drive a long-run average cost to a given value. The underlying motivation can be traced back to implementation issues associated with constrained optimal strategies.

A version of the ordinary differential equation (ODE) method as given by Metivier and Priouret is developed for proving almost sure convergence of this algorithm. This is done by exploiting the recurrence structure of the system under nonidling policies. A probabilistic representation of solutions to an associated Poisson equation is found most useful for proving their requisite Lipschitz continuity. The conditions that guarantee convergence are given directly in terms of the model data. The approach is of independent interest, as it is not limited to this particular queueing application and suggests a way of attacking other similar problems.

Key words. stochastic approximations, stochastic adaptive control, queueing networks

AMS(MOS) subject classifications. 90B22, 90B50, 93E20

1. Introduction.

1.1. Stochastic approximations on Markov chains. In recent years, there has been widespread interest in stochastic approximation algorithms as a means to solve increasingly complex engineering problems [5], [16]. As a result, focus has shifted from the original Robbins–Monro algorithm to *projected* stochastic approximation algorithms driven by *Markovian* “noise” or “state” processes. These algorithms have the following form: The state process $\{X(n), n = 0, 1, \dots\}$ takes values in some Borel subset S of \mathbb{R}^K . With U a compact convex subset of \mathbb{R}^p , the iterates $\{\eta(n), n = 0, 1, \dots\}$ are then defined by the recursion

$$(1.1) \quad \eta(0) \in U, \quad \eta(n+1) = \Pi_U \left\{ \eta(n) + a_{n+1} f(\eta(n), X(n+1)) \right\} \quad n = 0, 1, \dots,$$

where Π_U denotes the nearest-point projection on U , f is a Borel mapping $U \times S \rightarrow \mathbb{R}^p$ and the step size sequence $\{a_{n+1}, n = 0, 1, \dots\}$ satisfies some conditions, say (2.2) typically. For the Markovian dependencies alluded to earlier, a complete specification

* Received by the editors August 2, 1989; accepted for publication (in revised form) August 8, 1991.

† Electrical Engineering Department and Systems Research Center, University of Maryland, College Park, Maryland 20742. The work of this author was supported partially through Office of Naval Research grant N00014-84-K-0614, National Science Foundation grant ECS-83-51836, and United States–Israel Binational Science Foundation grant BSF 85-00306. Support provided by the Systems Research Center was made possible through the National Science Foundation’s Engineering Research Centers Program NSFD CDR-88-03012

‡ Electrical Engineering Department, Technion–Israel Institute of Technology, Haifa 32000, Israel. The work of this author was supported partially through United States–Israel Binational Science Foundation Grant BSF 85-00306 and partially through a grant from AT&T Bell Laboratories

of the algorithms (1.1) requires

$$(1.2) \quad P[X(n+1) \in B | X(0), \eta(0), X(1), \dots, X(n), \eta(n)] = \mu_{\eta(n)}(X(n); B)$$

$$n = 0, 1, \dots$$

for every Borel subset B of S , where $\{\mu_\eta, \eta \in U\}$ is a family of one-step probability transition kernels on S .

The central question in the theory of stochastic approximations is concerned with the convergence properties of the iterate sequence $\{\eta(n), n = 0, 1, \dots\}$. For the classical Robbins–Monro algorithm, Gladyshev [10] has given direct martingale arguments to establish almost sure convergence. However, in more complex situations such as (1.2), this direct probabilistic approach does not work, and this failure has prompted the development of the so-called ordinary differential equation (ODE) method. In most of its forms, the ODE method proceeds in two separate steps. The first step relies on the Kushner–Clark Lemma to identify a deterministic ODE, the stability properties of which determine the limit points of $\{\eta(n), n = 0, 1, \dots\}$. The second step is probabilistic in nature and depends on the algorithm being considered; its purpose is to show that asymptotically (in the mode of convergence of interest) the output sequence of the original algorithm behaves like the solution to the ODE.

In their monograph [17], Kushner and Clark have given general conditions for successfully completing this second step. In more structured situations [16], Kushner has shown how weak convergence methods pave the way to convergence in probability of the sequence $\{\eta(n), n = 0, 1, \dots\}$. In the Markovian case, Metivier and Priouret [23] have established almost sure convergence by making use of properties of the Poisson equation associated with the transition kernels $\{\mu_\eta, \eta \in U\}$ appearing in (1.2). Key to their analysis are various properties of Lipschitz continuity (in η) of the solution to this Poisson equation.

Unfortunately, in all these references, the conditions underlying the second step of the ODE method are given in implicit form and are often hard to verify in specific situations. What seems desirable is a more operational convergence theory where conditions are given *directly* in terms of the model data. This was done by the authors in the Markovian situation [18] when the state space S is *finite*. Under the mild condition of Lipschitz continuity (in η) for the one-step transition probabilities, almost sure convergence was established by a variant of the approach proposed by Metivier and Priouret.

When the state space S is *countably infinite*, the situation is much more difficult and no general results seem available, which guarantees almost sure convergence in terms of *explicit* conditions on the model data. The main technical difficulty in the approach of Metivier and Priouret stems from the fact that several quantities of interest are no longer bounded and that the requisite properties of the solution to the Poisson equation are now much harder to obtain. This paper presents arguments for establishing both these smoothness properties and the almost sure convergence of the algorithm. The general framework of interest is described in §2, and is couched in the formalism of the theory of Markov decision processes (MDPs); this is done for notational convenience as will become apparent in later sections. The approach advocated here relies on the recurrence structure of the (controlled) system [20], and on a probabilistic representation of the solution to the Poisson equation derived from it [30]. These arguments are developed in the context of an adaptive control problem

for a specific queueing system, namely, a discrete-time single-server network with random routing, which is described in §3. The approach presented here is of much wider applicability and should be of use in analyzing a large class of projected stochastic approximations driven by a Markov chain on a countable state space. The main advantage of discussing a concrete application lies in the fact that the key arguments can then be provided in their simplest form, unencumbered by often confusing technicalities, under *verifiable* conditions given solely in terms of the model data. To help the reader apply the ideas proposed here to other situations, each one of §§5–7 ends with an outline of more general technical conditions, which permits a development similar to the one given here.

1.2. A time-sharing queueing system. The queueing system considered here is now briefly described; a precise model formulation is available in §3: Consider a system composed of K infinite capacity queues that compete for the use of a single server. Time is slotted with the service requirement of each customer corresponding exactly to one time slot. At the beginning of each time slot, the controller gives priority to one of the queues according to some prespecified dynamic priority assignment, and the selected queue is given service attention during that slot. However, due to a variety of reasons ranging from server failure to exogenous interferences, with a positive probability, the service fails, in which case the service of that customer is rescheduled at a later time in accordance with the service allocation policy. When in a given time slot the service succeeds, the customer is either declared serviced and leaves the system at the end of the slot or is routed to one of the other queues with a fixed probability, depending on both source and destination queues. The failures are assumed generated through *independent Bernoulli* processes, with possibly class-dependent parameters, and this *independently* of the arrival mechanism. New customers may arrive in batches which are modeled as an arbitrary K -dimensional *renewal* process; this captures possible partial correlations between arrivals from different classes in a given slot.

This queueing system and its variants constitute useful models for studying issues of resource allocation in several application areas, including computer systems and data networks, and as such they have received a great deal of attention in recent years. Klimov [14] studied a continuous-time version of this system and proved that a strict priority policy minimizes the discounted cost associated with a cost-per-slot linear in the queue sizes. Tsoucas and Walrand [31] considered an adaptive version of Klimov's problem where the service distributions are unknown.

The case where no routing is allowed has been much studied: Several authors [3], [4], [8], [11] have shown that the μc -rule minimizes a variety of performance measures associated with the aforementioned linear cost structure. In [24], Nain and Ross considered the situation where several types of traffic, say voice, video, and data, compete for the use of a single synchronous communication channel. They formulated this situation as a system of K discrete-time queues and found the service allocation strategy that minimizes the long-run average of a linear expression in the queue sizes of $K - 1$ customer classes, under the constraint that the long-run average queue size of the remaining customer class not exceed a certain value. Extending some of the optimality results from Baras, Ma, and Makowski [4], they showed that if the constraint can be met, then the optimal policy g is a Markov stationary policy with the following structure: There exist two *static* work-conserving service assignment policies (of which μc -rules are only one description), say \bar{g} and \underline{g} , and a scalar η^* in $(0, 1)$. At the beginning of each time slot, a coin with bias η^* is flipped, and the policy

g implements channel rights according to the outcome via \bar{g} and \underline{g} with probability η^* and $1 - \eta^*$, respectively. The bias η^* is determined so as to meet the constraint. This result was extended by Altman and Shwartz to the case where the constraint is also given through a linear combination in the queue sizes [1], [2].

These results are typical in the broader context of MDPs in that analysis often identifies a policy g of interest which is Markov stationary. In fact, for the problem of minimizing one average cost subject to a constraint on another such cost, an optimal policy which “mixes” two deterministic policies in the manner described above exists under very general conditions [6], [7], [27]. Unfortunately, this policy may not be readily implementable due either to a lack of knowledge of the actual values of some parameters [15] or to computational difficulties inherent to its definition. The situation treated by Nain and Ross [24] is a good case in point, for there nontrivial off-line computations are required to actually compute the value of the bias η^* , even if all parameters are known.

1.3. Overview of the paper. This implementation issue provides the motivation for the stochastic approximation studied in this paper. In §4, the issue is discussed in the broader context of “steering the cost to a given value” [19], with a view towards applications to constrained optimization [1], [2], [26]. The problem is now one of finding the bias η^* needed in a simple randomization between two policies \underline{g} and \bar{g} to steer a long-run average cost to a given value. The resulting randomized Markov stationary policy—denoted g hereafter—can be implemented by means of a projected stochastic approximation. This algorithm computes on-line estimates of η^* which are then used in a certainty equivalence controller α derived from the special form of g . Theorems 4.1 and 4.2 contain the main results concerning the performance of this policy α , namely, that the policies α and g yield the same value for the long-run average cost, and that under α the iterates $\{\eta(n), n = 0, 1, \dots\}$ converge almost surely to the bias value η^* . This improves on earlier results of the authors [28] for the same algorithm in the context of the two-queue system with no routing. There, only convergence in probability was established, albeit under weaker conditions on moments.

The convergence proof for the stochastic approximation algorithm hinges on the availability of bounds on moments of the queue size process which are uniform in the policy, and on the smoothness properties of solutions to an associated Poisson equation [23], [30]. The bounds are obtained in §5 by means of renewal arguments that relate the queue size to the recurrence times to the empty state. In §6, novel arguments are developed for proving the Lipschitz continuity of solutions to the Poisson equation and for establishing bounds on them. It is appropriate to stress the methodological value of both §§5 and 6, in that ideas therein are by no means restricted to the competing queue model or to the randomization of two policies, and can be used *mutatis mutandis* in many other situations. However, the approach was developed here in the context of a specific model, rather than for general Markov chains with countable state spaces, to present the arguments more clearly, unencumbered by technical details and assumptions which often accompany more formal treatments.

The almost sure convergence of the stochastic approximation scheme defining the implementation α is established in §7, where the various estimates of the previous sections allow for a rather simple proof. Finally, the cost properties of the policy α are discussed in §8 by making use of the convergence of the stochastic approximation and by invoking the results on the certainty equivalence principle developed in [30]; the requisite hypotheses of [30] are easily verified for this system with the help of bounds

on solutions to the Poisson equation. The paper concludes with an application to the constrained optimization problem discussed by Nain and Ross in [24]. All necessary conditions are verified and the policy α thus constitutes an implementation of the Markov stationary policy which is constrained optimal for this problem.

2. A general model. This section introduces a general class of projected stochastic approximations driven by Markovian noise. The formalism of the theory of MDPs [12], [25] was found notationally convenient as it lends itself naturally to the presentation of the more general conditions at the end of §§5–7.

A few words on the notation and conventions used throughout the paper. The set of all nonnegative integers is denoted by \mathbb{N} , and \mathbb{R} (respectively, \mathbb{R}_+) stands for the set of all real (respectively, positive real) numbers. The indicator function of a set A is denoted by $I[A]$. Unless stated otherwise, the notation \lim_n and $\overline{\lim}_n$ are understood with n going to infinity. The infimum over an empty set is taken to be ∞ .

2.1. The MDP formulation. Consider an MDP (S, U, P) as defined in the literature [12], [25] where the state space S is a countable set and the action space U is a compact convex subset of \mathbb{R}^p . The one-step transition mechanism P is defined through the one-step transition probability functions $U \rightarrow [0, 1] : u \rightarrow p_{xy}(u)$ (with x, y in S), which are assumed Borel measurable and which satisfy $\sum_y p_{xy}(u) = 1$ for all u in U , and all x in S . The space of probability measures on U (when equipped with its natural Borel σ -field) is denoted by $\mathcal{M}(U)$. An *admissible* control policy π is then defined as any collection $\{\pi_n, n = 0, 1, \dots\}$ of mappings $\pi_n : S \times (U \times S)^n \rightarrow \mathcal{M}(U)$ such that for all $n = 0, 1, \dots$ and every Borel subset B of U , the mapping $S \times (U \times S)^n \rightarrow [0, 1] : h_n \rightarrow \pi_n(h_n; B)$ is Borel measurable.

The definition of the MDP (S, U, P) postulates the existence of a measurable space (Ω, \mathcal{F}) large enough to carry sequences of S -valued random variables (rvs) $\{X(n), n = 0, 1, \dots\}$ and U -valued rvs $\{U(n), n = 0, 1, \dots\}$, with $X(n)$ denoting the state of the system at time n and $U(n)$ representing the action taken in that state. The feedback information is encoded through the rvs $\{H(n), n = 0, 1, \dots\}$ defined by $H(0) := X(0)$ and $H(n) := (X(0), U(0), X(1), \dots, U(n-1), X(n))$ for all $n = 1, 2, \dots$; the rv $H(n)$ takes values in $\mathbb{H}_n := S \times (U \times S)^n$. The measurable space (Ω, \mathcal{F}) is often selected to be the so-called canonical space, i.e., Ω is the Cartesian product $\Omega := S \times (U \times S)^\infty$ endowed with the natural Borel structure inherited from the product topology. However, in many concrete situations, it is more convenient to describe the underlying MDP on a measurable space (Ω, \mathcal{F}) which is somewhat larger than the canonical space. For example, for the network considered in this paper, additional rvs are needed to encode arrivals, service completions and random routing in the queueing system, in which case the definitions of \mathbb{H}_n and $H(n)$ are modified accordingly in the obvious way. To complete the definition of the MDP (S, U, P) , let $\mu(\cdot)$ be a fixed probability distribution on S . For every admissible policy π , a probability measure P^π is constructed on (Ω, \mathcal{F}) in the usual way [12], [25] such that under P^π , the rv X_0 has distribution $\mu(\cdot)$. The expectation operator associated with π (or P^π) is denoted by E^π .

Following standard usage, an admissible policy π is said to be a *Markov* or *memoryless* policy if there exists a family $\{g_n, n = 0, 1, \dots\}$ of Borel mappings $g_n : S \rightarrow \mathcal{M}(U)$ such that $\pi_n(\cdot; H(n)) = g_n(\cdot; X(n))$ P^π -almost surely for all $n = 0, 1, \dots$. When the mappings $\{g_n, n = 0, 1, \dots\}$ are all identical to a given mapping $g : S \rightarrow \mathcal{M}(U)$, the Markov policy is termed *stationary* and is identi-

fied with the mapping g itself. Under any Markov stationary g , the state process $\{X(n), n = 0, 1, \dots\}$ evolves according to a Markov chain. Finally, an admissible policy π is said to be deterministic or nonrandomized if there exists a sequence of Borel mappings $\{f_n, n = 0, 1, \dots\}$ such that for each $n = 0, 1, \dots$, the mapping $f_n : \mathbb{H}_n \rightarrow U$ is Borel measurable and the probability measure $\pi_n(\cdot; H(n))$ is a point mass distribution concentrated at $f_n(H(n))$ P^π -almost surely.

2.2. The stochastic approximation. Stochastic approximations on Markov chains—as defined by (1.1) and (1.3)—can be interpreted as deterministic policies for the MDP (S, U, P) described earlier. To see this, start with a mapping $c : S \rightarrow \mathbb{R}^p$ and let $\{\eta(n), n = 0, 1, \dots\}$ be the sequence of U -valued rvs determined by the recursion

$$(2.1) \quad \eta(0) \in U, \quad \eta(n+1) = \Pi_U \{ \eta(n) + a_{n+1} c(X(n+1)) \}, \quad n = 0, 1, \dots$$

As before, Π_U denotes the nearest-point projection on U , and the stepsize sequence $\{a_{n+1}, n = 0, 1, \dots\}$ satisfies the usual conditions

$$(2.2) \quad 0 < a_n \downarrow 0, \quad \sum_{n=0}^{\infty} a_n = \infty \quad \text{and} \quad \sum_{n=0}^{\infty} a_n^2 < \infty.$$

The policy associated with the recursion (2.1) is the deterministic policy $\alpha = \{\alpha_n, n = 0, 1, \dots\}$ with the property that for all $n = 0, 1, \dots$, $\alpha_n(\cdot; H(n))$ is the point mass distribution concentrated at $\eta(n)$. This policy is admissible since for each $n = 0, 1, \dots$, the rv $\eta(n)$ can be expressed as a function of the successive states $X(0), X(1), \dots, X(n)$.

3. The discrete-time Klimov model. This section presents in some detail the model for the controlled queueing system briefly described in the introduction. First, a few words on the notation and convention in use. Elements of \mathbb{R}^K are always interpreted as $K \times 1$ column vectors, and the k th component of any element x of \mathbb{R}^K is denoted by $x_k, k = 1, \dots, K$, with a similar convention for rvs. Thus an element x of \mathbb{R}^K can also be written as $(x_1, \dots, x_K)'$ (with $'$ denoting transpose), and its norm is given by $|x| := \sum_{k=1}^K |x_k|$. The standard basis $\{e^1, \dots, e^K\}$ for \mathbb{R}^K is denoted by \mathcal{B}_K , while the standard simplex \mathcal{S}_K is defined by $\mathcal{S}_K := \{p \in [0, 1]^K : \sum_{k=1}^K p_k = 1\}$; it is plain that \mathcal{S}_K can be identified with $\mathcal{M}(U)$ when $U = \mathcal{B}_K$.

3.1. The basic random variables. The controlled queueing system of interest, the so-called discrete-time Klimov model, is defined as an MDP with all its probabilistic elements defined on a single sample space Ω equipped with the σ -field of events \mathcal{F} . This sample space carries the basic rvs $\Xi, \{U(n), n = 0, 1, \dots\}, \{A(n), n = 0, 1, \dots\}, \{B(n), n = 0, 1, \dots\}$ and $\{R(n), n = 0, 1, \dots\}$ which take values in $\mathbb{N}^K, \mathcal{B}_K, \mathbb{N}^K, \{0, 1\}^K$, and $\{0, 1, \dots, K\}^K$, respectively. These quantities have a ready interpretation in the context of the queueing system described in the introduction: For $k = 1, \dots, K$, the number of customers initially in the k th queue is set at Ξ_k and for each $n = 0, 1, \dots$, the state of the system is represented by a \mathbb{N}^K -valued rv $X(n)$ with the interpretation that at the beginning of the slot $[n, n+1)$, $X_k(n)$ customers are present in the k th queue, including the one receiving service. The following chain of events then occurs:

(i) The control action $U(n)$ is selected with the convention that $U_k(n) = 1$ (respectively, $U_k(n) = 0$) if the k th queue is (respectively, is not) given service attention during that slot. The fact that $U(n)$ takes values in \mathcal{B}_K guarantees that exactly one queue is given service attention;

(ii) New customers arrive into the system according to the rv $A(n)$ with $A_k(n)$ new customers joining the k th queue;

(iii) A completion of service possibly occurs at the queue that was given service attention during the slot. This is encoded in the binary rv $B(n)$, where $B_k(n) = 1$ (respectively, $B_k(n) = 0$) signifies successful completion (respectively, abortion) of service for the k th queue conditioned on it being given service attention and nonempty; and

(iv) If a service completion occurs at the queue that was given service attention during the slot, then instantaneously the serviced customer is either transferred to another queue or it leaves the network. This routing decision is implemented through the variable $R(n)$ with the following interpretation. If the service completion occurred at the k th queue, then $R_k(n) = \ell$, $\ell = 1, \dots, K$, means that the serviced customer joins the ℓ th queue while $R_k(n) = 0$ expresses the fact that this customer leaves the system.

As a result of (i)–(iv), the successive system states or queue size vectors form a sequence $\{X(n), n = 0, 1, \dots\}$ of \mathbb{N}^K -valued rvs which are generated componentwise through the recursion

$$(3.1) \quad \begin{aligned} X_k(0) &= \Xi_k, \quad X_k(n+1) = X_k(n) + A_k(n) - I[X_k(n) \neq 0]U_k(n)B_k(n) \\ &\quad + \sum_{\ell=1}^K I[X_\ell(n) \neq 0]U_\ell(n)B_\ell(n)I[R_\ell(n) = k], \\ &\quad k = 1, \dots, K, \quad n = 0, 1, \dots \end{aligned}$$

At the beginning of each time slot $[n, n+1)$, the decision-maker has knowledge of the rv $H(n)$ which here includes the initial queue sizes, past arrivals, past decisions, past service completions, and past routing decisions so that the rvs $\{H(n), n = 0, 1, \dots\}$ are now given recursively by

$$(3.2) \quad H(0) = \Xi, \quad H(n+1) := (H(n), U(n), A(n), B(n), R(n)), \quad n = 0, 1, \dots$$

The information contained in $H(n)$ is used to generate the control value $U(n)$ implemented in the slot $[n, n+1)$.

3.2. The probabilistic structure. Since randomized strategies are allowed, an admissible control policy π is defined as any collection $\{\pi_n, n = 0, 1, \dots\}$ of mappings $\pi_n : \mathbb{H}_n \rightarrow \mathcal{S}_K$, with the interpretation that at times $n = 0, 1, \dots$, the k th queue is given service attention with probability $\pi_n(k; h_n)$ whenever the information vector h_n (in \mathbb{H}_n) is available to the system controller. The collection of all such admissible policies is denoted by \mathcal{P} . A policy π in \mathcal{P} is said to be *nonidling* whenever conditions $[\pi_n(k; H(n)) > 0, X(n) \neq 0] = [\pi_n(k; H(n)) > 0, X_k(n) \neq 0]$, $k = 1, \dots, K$, hold true P^π -almost surely for all $n = 0, 1, \dots$.

Let $q_\Xi(\cdot)$ and $q(\cdot)$ be two probability mass distributions on \mathbb{N}^K , and fix a service rate vector μ in $(0, 1]^K$. Moreover, let $P \equiv (p_{k\ell})$ denote a $K \times K$ substochastic matrix, i.e., $0 \leq p_{k\ell} \leq 1$ and $\sum_{\ell=1}^K p_{k\ell} \leq 1$ for all $k, \ell = 1, \dots, K$, and set

$$(3.3) \quad p_{k0} := 1 - \sum_{\ell=1}^K p_{k\ell}, \quad k = 1, \dots, K.$$

Throughout the discussion, the *nondegeneracy* and *finite mean* conditions

$$(3.4) \quad 0 < q(0) < 1 \quad \text{and} \quad \sum_{a \in \mathbb{N}^K} |a|q(a) < \infty$$

are enforced. Moreover, the matrix $I - P$ is assumed *invertible*, a condition which is equivalent to the system being open, i.e., every customer eventually leaves the system with probability one.

The model is now completely specified by postulating the existence of a family $\{P^\pi, \pi \in \mathcal{P}\}$ of probability measures on the σ -field \mathcal{F} which satisfies the requirements **(R1)**–**(R3)** below, i.e., for every policy π in \mathcal{P} ,

(R1) The rv Ξ is distributed according to $q_\Xi(\cdot)$, i.e., $P^\pi[\Xi = x] = q_\Xi(x)$ for all x in \mathbf{N}^K ;

(R2) For all a in \mathbf{N}^K , b in $\{0, 1\}^K$ and r in $\{0, 1, \dots, K\}^K$,

$$P^\pi[A(n) = a, B(n) = b, R(n) = r \mid \mathcal{F}_n \vee \sigma\{U(n)\}] \\ = q(a) \cdot \prod_{k=1}^K (b_k \mu_k + (1 - b_k)(1 - \mu_k)) \cdot \prod_{k=1}^K p_{kr_k}, \quad n = 0, 1, \dots,$$

where $\mathcal{F}_n = \sigma\{H(n)\}$ with $H(n)$ defined by (3.2); and

(R3) For all $k = 1, \dots, K$,

$$P^\pi[U(n) = e^k \mid \mathcal{F}_n] := \pi_n(k; H_n), \quad n = 0, 1, \dots$$

A sample space (Ω, \mathcal{F}) that carries such a family of probability measures $\{P^\pi, \pi \in \mathcal{P}\}$ is easily constructed by taking Ω to be the canonical space $\mathbf{N}^K \times (\mathcal{B}_K \times \mathbf{N}^K \times \{0, 1\}^K \times \{0, 1, \dots, K\}^K)^\infty$ equipped with its natural σ -field; the reader is referred to [20], [21], [28] for additional details on this construction. The basic rvs satisfy various independence and distributional properties referred to as properties **(P1)**–**(P4)**, i.e., under P^π , for each policy π in \mathcal{P} ,

(P1) The \mathbf{N}^K -valued rv Ξ and the sequences of rvs $\{A(n), n = 0, 1, \dots\}$, $\{B(n), n = 0, 1, \dots\}$, and $\{R(n), n = 0, 1, \dots\}$ are *mutually independent*;

(P2) The \mathbf{N}^K -valued rvs $\{A(n), n = 0, 1, \dots\}$ form a sequence of *independent and identically distributed* (i.i.d.) rvs with common probability distribution $q(\cdot)$;

(P3) The sequences $\{B_k(n), n = 0, 1, \dots\}$ of $\{0, 1\}$ -valued rvs are *mutually independent i.i.d. Bernoulli* sequences with parameters $\mu_k, k = 1, \dots, K$; and

(P4) The sequences $\{R_k(n), n = 0, 1, \dots\}$ of $\{0, 1, \dots, K\}$ -valued rvs are *mutually i.i.d.* sequences with $P^\pi[R_k(n) = \ell] = p_{k\ell}, k, \ell = 1, \dots, K$, for all $n = 0, 1, \dots$.

For $k = 1, \dots, K$, denote by λ_k the first moment of the sequence $\{A_k(n), n = 0, 1, \dots\}$ and set $\nu_k = \mu_k^{-1}$. The *network traffic* coefficient ρ is then defined by

$$(3.5) \quad \rho := \lambda'(I - P)^{-1}\nu,$$

where $\lambda = (\lambda_1, \dots, \lambda_K)'$ and $\nu = (\nu_1, \dots, \nu_K)'$.

4. Problem formulation. For any mapping $c : \mathbf{N}^K \rightarrow \mathbb{R}$, set

$$(4.1) \quad J_c(\pi) := \overline{\lim}_n E^\pi \left[\frac{1}{n+1} \sum_{i=0}^n c(X(i)) \right], \quad \pi \in \mathcal{P}$$

(whenever meaningful) with the usual interpretation that $J_c(\pi)$ is a measure of system performance when using the policy π .

4.1. Steering the cost. Constrained MDPs lead to optimal stationary policies that randomize between several stationary deterministic policies [6], [7], [26], [27]. Given the constituent deterministic policies, the problem of finding the optimal policy reduces to simultaneously *steering* constraint functionals of the form (4.1) to given values. For simplicity, only the scalar case (arising from a single constraint) is discussed here, in which case the steering problem consists in finding a Markov stationary policy g such that $J_c(g) = V$ for some given constant V . The discussion is given under the assumption that there exist two Markov (possibly randomized) stationary policies \underline{g} and \bar{g} such that

$$(4.2) \quad J_c(\underline{g}) < V < J_c(\bar{g}).$$

For every η in $[0,1]$, let f^η denote the Markov stationary policy obtained by simply randomizing with *bias* η between the policies \underline{g} and \bar{g} ; it is determined through the mapping $f^\eta : \mathbf{N}^K \rightarrow \mathcal{S}_K$, where

$$(4.3) \quad f^\eta(k; x) := \eta \bar{g}(k; x) + (1 - \eta) \underline{g}(k; x), \quad x \in \mathbf{N}^K, \quad k = 1, \dots, K.$$

For $\eta = 1$ (respectively, $\eta = 0$) the randomized policy f^η coincides with \bar{g} (respectively, \underline{g}). If the mapping $\eta \rightarrow J_c(f^\eta)$ is *continuous* on the interval $[0,1]$, then by virtue of (4.2) at least one randomized strategy f^{η^*} meets the value V and its corresponding bias value η^* is a solution to the equation

$$(4.4) \quad J_c(f^{\eta^*}) = V, \quad \eta^* \in [0, 1],$$

so that the identification $g = f^{\eta^*}$ may take place.

4.2. Implementation issues. Solving the (highly) nonlinear equation (4.4) for the bias value η^* is usually a nontrivial task, even in the simplest of situations [19], [24]. This difficulty is circumvented by proposing alternatives to the policy g that bypass a *direct* solution of (4.4). One possible approach is to design (simple recursive) schemes for estimating the value η^* which solves (4.4) and then to define a so-called “naive feedback” policy $\alpha = \{\alpha_n, n = 0, 1, \dots\}$ via the certainty equivalence principle [22]. Such a policy α can be written in the form

$$(4.5) \quad \alpha_n := \eta(n) \bar{g} + (1 - \eta(n)) \underline{g}, \quad n = 0, 1, \dots$$

for some sequence of $[0,1]$ -valued rvs $\{\eta(n), n = 0, 1, \dots\}$ which act as “estimates” for the bias value η^* . It is hoped that the effects of controlling and learning about the system will combine to produce a *consistent* estimation scheme. In such a case, the sequence of estimates $\{\eta(n), n = 0, 1, \dots\}$ converges to the value η^* in some sense, thus providing increasingly better approximations to the appropriate bias value. This policy α will constitute an acceptable *implementation* of g provided $J_c(\alpha) = J_c(g)$.

At this point, the reader may wonder as to how such an estimation scheme can be selected. If the function $\eta \rightarrow J_c(f^\eta)$ were *continuous* and *strictly monotone* (necessarily increasing by (4.2)–(4.3)), then the search for η^* could be interpreted as finding the zero of the continuous, strictly monotone function $\eta \rightarrow J_c(f^\eta) - V$, and this brings to mind ideas from the theory of *stochastic approximations* [17]. Here, the Robbins–Monro version of these algorithms suggests that a sequence of bias values $\{\eta(n), n = 0, 1, \dots\}$ be generated through the recursion

$$(4.6) \quad \eta(0) \in U, \quad \eta(n+1) = \left[\eta(n) + a_{n+1}(V - c(X(n+1))) \right]_0^1, \quad n = 0, 1, \dots$$

In (4.6) the notation $[x]_0^1 = 0 \vee (x \wedge 1)$ is used for every x in \mathbb{R} , and the sequence of stepsizes $\{a_n, n = 1, 2, \dots\}$ satisfies the conditions (2.2).

4.3. The results. This paper is devoted to analyzing the performance of the adaptive policy α defined through (4.5)–(4.6). The main results, which are described below, require the additional assumptions **(R4)**–**(R6)** on the data of the problem, where

(R4) There exists some integer $\gamma \geq 1$ such that the moment conditions

$$\sum_{x \in \mathbf{N}^K} |x|^\gamma q_\Xi(x) < \infty \quad \text{and} \quad \sum_{a \in \mathbf{N}^K} |a|^\gamma q(a) < \infty$$

hold true, i.e., under every policy π in \mathcal{P} , $E^\pi[|\Xi|^\gamma] < \infty$ and $E^\pi[|A(n)|^\gamma] < \infty$ for all $n = 0, 1, \dots$;

(R5) There exist an integer $\delta > 0$ and a constant $L > 0$ such that

$$|c(x)| \leq L(1 + |x|^\delta) =: \tilde{c}(|x|), \quad x \in \mathbf{N}^K;$$

(R6) The policies \bar{g} and \underline{g} are *non-idling* Markov stationary policies such that (4.2) holds.

THEOREM 4.1. Assume **(R1)**–**(R6)** to hold with $\rho < 1$ and let the integer exponent γ in **(R4)** and δ in **(R5)** satisfy the condition

$$(4.7) \quad 2\delta + 3 \leq \gamma.$$

If the mapping $\eta \rightarrow J_c(f^\eta)$ is strictly monotone, then $\lim_n \eta(n) = \eta^*$ P^α -almost surely.

Under these conditions, the system also satisfies a certainty equivalence principle [29].

THEOREM 4.2. Assume **(R1)**–**(R6)** to hold with $\rho < 1$ and let the integer exponent γ in **(R4)** and δ in **(R5)** satisfy the condition

$$(4.8) \quad \max\{3, 1 + \delta(1 + \epsilon)\} \leq \gamma$$

for some $\epsilon > 0$. If $\lim_n \eta(n) = \eta^*$ in probability under P^α , then the convergence

$$(4.9) \quad J_c(\alpha) = \lim_n \frac{1}{n+1} \sum_{i=0}^n c(X(i)) = J_c(g)$$

takes place in $L^1(\Omega, \mathcal{F}, P^\alpha)$, so that

$$(4.10) \quad J_c(\alpha) = \lim_n E^\alpha \left[\frac{1}{n+1} \sum_{i=0}^n c(X(i)) \right] = J_c(g).$$

Moreover, for any other mapping $d: \mathbf{N}^K \rightarrow \mathbb{R}$, if there exist an integer $\delta' > 0$ and a constant $L' > 0$ such that

$$(4.11) \quad |d(x)| \leq L'(1 + |x|^{\delta'}), \quad x \in \mathbf{N}^K$$

then both (4.9) and (4.10) hold for the long-run average cost (4.1) associated with d provided the condition

$$(4.12) \quad \max\{3, 1 + \delta'(1 + \epsilon')\} \leq \gamma$$

holds for some $\epsilon' > 0$.

This section closes with a few easy facts and remarks: The restriction that δ and δ' be integers in Theorems 4.1 and 4.2 is not essential but results in some simplifications in the notation. An example where the hypotheses of Theorems 4.1 and 4.2 hold is given in §8.

Under **(R6)**, the policies $f^\eta, 0 \leq \eta \leq 1$ and α are all nonidling since \bar{g} and \underline{g} are nonidling.

For each $n = 0, 1, \dots$, (3.1) implies $X_k(n+1) \leq X_k(n) + A_k(n) + 1, k = 1, \dots, K$, whence by virtue of **(R4)**, $E^\pi[|X(n)|^\gamma] < \infty$ under any policy π in \mathcal{P} . Since $\delta \leq \gamma$ under either (4.7) or (4.8), it is then immediate from **(R5)** that $E^\pi[|c(X(n))|] < L(1 + E^\pi[|X(n)|^\delta]) < \infty$, and therefore $J_c(\pi)$ is always well defined (and in fact finite by Theorem 5.1 below). A similar argument shows that under the conditions (4.11)–(4.12), the long-run average cost associated with d is also well defined and finite under any policy π in \mathcal{P} .

5. Moment estimates.

5.1. The bounds. The proofs of Theorems 4.1 and 4.2 require bounds on moments of the rvs $\{|X(n)|, n = 0, 1, \dots\}$ which are uniform over the class of *all* nonidling policies. The derivation of such bounds is given below and is based on the key observation that the *total* number of customers in the system at any given time n decreases by *at most one* unit in the next time slot $[n, n+1)$, and is therefore bounded above by the number of slots it takes for the queue sizes to empty for the first time after n . This simple fact can be used to advantage when combined with the detailed statistical information obtained by the authors in [20] on the time until the system empties, and leads to the following strong estimates.

THEOREM 5.1. *Assume **(R1)**–**(R5)** with $\rho < 1$. There exists a single positive constant K_γ such that for every nonidling policy π in \mathcal{P} , the moment estimate*

$$(5.1) \quad \sup_n E^\pi[|X(n)|^{\gamma-1}] \leq K_\gamma < \infty$$

holds true.

Theorem 5.1, the proof of which is presented below, turns out to be a special case of an intermediate result of independent interest given in Theorem 5.4. Before discussing this more general result, it is convenient to note the following simple and useful consequence of (5.1).

COROLLARY 5.2. *Assume **(R1)**–**(R5)** with $\rho < 1$. Whenever $\gamma > 2$, the rvs $\{|X(n)|, n = 0, 1, \dots\}$ are uniformly integrable under the probability measure P^π associated with any non-idling policy π in \mathcal{P} .*

5.2. Recurrence properties. To formalize the argument outlined earlier, it is necessary to study the recurrence structure of the process $\{X(n), n = 0, 1, \dots\}$ under any nonidling policy π in \mathcal{P} . To that end, consider the rvs $\{\tau_k, k = 0, 1, 2, \dots\}$ and $\{\sigma_k, k = 1, 2, \dots\}$ defined recursively, with $\tau_0 = \sigma_1 := 0$, by

$$(5.2a) \quad \tau_k := \inf\{n > \sigma_k : X(n) = 0\}, \quad k = 1, 2, \dots$$

and

$$(5.2b) \quad \sigma_{k+1} := \inf\{n > \tau_k : X(n) \neq 0\}, \quad k = 1, 2, \dots$$

These definitions are different from those given in [20] (where $\nu(0)$ denotes the present rv τ_1). For $k = 2, 3, \dots$ the rv τ_k is the time epoch at which the system empties itself

for the $(k-1)$ st time after τ_1 , so that σ_{k+1} is the time epoch when the system becomes again nonempty for the first time after τ_k . Moreover, with the convention $\infty - \infty = 0$, define the rvs $\{\theta_k, k = 1, 2, \dots\}$ by

$$(5.3) \quad \theta_{k+1} = \tau_{k+1} - \tau_k, \quad k = 0, 1, \dots$$

so that $\theta_1 = \tau_1$. The following proposition summarizes results that were obtained in [20, §§4 and 5].

PROPOSITION 5.3. *Assume (R1)–(R4) with $\rho < 1$. Under any nonidling policy π in \mathcal{P} , the rvs $\{\theta_k, k = 1, 2, \dots\}$ form a delayed renewal process whose statistics are independent of the policy π , with finite means given by*

$$(5.4) \quad E^\pi[\theta_k | X(0) = x] = \begin{cases} \frac{1}{1-\rho} \cdot x'(I-P)^{-1}\nu + \frac{1}{1-\rho}I[x=0] & \text{if } k = 1 \\ \frac{1}{1-q(0)} \cdot \frac{1}{1-\rho} & \text{if } k = 2, 3, \dots \end{cases}$$

for all x in \mathbb{N}^K . Moreover, the rv θ_2 possesses finite moments of order γ , and for $\ell = 1, \dots, \gamma$, there exists a positive constant C_ℓ (independent of the policy π) such that

$$(5.5) \quad E^\pi[\tau_1^\ell | X(0) = x] \leq C_\ell(1 + |x|^\ell), \quad x \in \mathbb{N}^K.$$

In view of this result, it is natural to introduce \mathcal{E}_x as the expectation operator with respect to the distribution of τ_1 given that $X(0) = x$ and that *any* nonidling policy is used. Finally, for reference, denote by G the distribution of the rv $\theta_1 (= \tau_1)$ and by F the common distribution of the i.i.d. rvs $\{\theta_k, k = 2, 3, \dots\}$. By definition, the distributions G and F do not coincide.

5.3. A renewal estimate. The (continuous-time) counting process $\{N(t), t \geq 0\}$ naturally associated with the sequence $\{\tau_n, n = 0, 1, \dots\}$ is defined by

$$(5.6) \quad N(t) := \max\{k = 0, 1, \dots : \tau_k \leq t\}, \quad t \geq 0$$

with the ready interpretation that $N(t)$ represents the number of times the queue has returned to the empty state by time t . With this notation, the observation made earlier translates into

$$(5.7) \quad |X(n)| \leq \tau_{N(n)+1} - n, \quad n = 0, 1, \dots$$

Now, for any *monotone nondecreasing* mapping $r : \mathbb{R}_+ \rightarrow \mathbb{R}_+$, set

$$(5.8) \quad R_G(t) := E_\Xi^\pi[r(\tau_{N(t)+1} - t)], \quad t \geq 0.$$

The subscripts G and Ξ in (5.8) emphasize the fact that the system is started with an initial queue size Ξ distributed according to the distribution q_Ξ . Since the sequence $\{\theta_k, k = 2, 3, \dots\}$ is a *nondelayed* renewal sequence, it is appropriate to define

$$(5.9) \quad R_F(t) := E^\pi[r(\tau_{N(t+\tau_1)+1} - (t + \tau_1))], \quad t \geq 0$$

as this corresponds to a nondelayed renewal process with $G = F$. The first part of this section is devoted to deriving a bound on the expected values $\{R_G(t), t \geq 0\}$ for any nonidling policy π , with a view toward generating (via (5.7)) a bound for the sequence of expected values $\{E^\pi[|X(n)|], n = 0, 1, \dots\}$.

THEOREM 5.4. *Assume **(R1)**–**(R4)** with $\rho < 1$ and let π be an arbitrary non-idling policy in \mathcal{P} . Under the finite moment assumptions*

$$(5.10) \quad m_G(r) := \int_0^\infty r(\theta) dG(\theta) < \infty \quad \text{and} \quad m_F(r) := \int_0^\infty r(\theta) dF(\theta) < \infty,$$

the condition

$$(5.11) \quad K_F(r) := \int_0^\infty \int_0^\theta r(\theta - t) dt dF(\theta) = \int_0^\infty \int_t^\infty r(\theta - t) dF(\theta) dt < \infty$$

implies

$$(5.12) \quad \sup_{t \geq 0} R_G(t) = \sup_{t \geq 0} E_\pi^\pi[r(\tau_{N(t)+1} - t)] < \infty.$$

Proof. Let r_G and r_F be the mappings $\mathbb{R}_+ \rightarrow \mathbb{R}_+$ defined by

$$(5.13) \quad r_G(t) := \int_t^\infty r(\theta - t) dG(\theta) \quad \text{and} \quad r_F(t) := \int_t^\infty r(\theta - t) dF(\theta), \quad t \geq 0.$$

The finiteness conditions (5.10) translate into $r_G(0) = m_G(r) < \infty$ and $r_F(0) = m_F(r) < \infty$. Since r takes positive values and is monotone nondecreasing, the indefinite integrals entering the definition (5.13) are well defined and satisfy the inequalities (5.14)

$$0 \leq \int_t^\infty r(\theta - t) dG(\theta) \leq \int_t^\infty r(\theta - s) dG(\theta) \leq \int_s^\infty r(\theta - s) dG(\theta), \quad 0 \leq s \leq t.$$

As a result, the mapping r_G is well defined and monotone nonincreasing. Similar comments hold for r_F .

A standard renewal argument [13, p. 183] applied to the process $\{r(\tau_{N(t)+1} - t), t \geq 0\}$ shows that

$$(5.15) \quad R_G(t) = \int_0^t R_F(t - \theta) dG(\theta) + \int_t^\infty r(\theta - t) dG(\theta), \quad t \geq 0$$

whence

$$(5.16) \quad \begin{aligned} R_G(t) &\leq \int_0^t R_F(t - \theta) dG(\theta) + \int_0^\infty r(\theta) dG(\theta) \\ &\leq \sup_{0 \leq s \leq t} R_F(s) + m_G(r), \quad t \geq 0 \end{aligned}$$

by the remarks made earlier. This clearly shows that under (5.10), the result (5.12) will hold if the bound

$$(5.17) \quad \sup_{t \geq 0} R_F(t) < \infty$$

can be established.

When $G = F$, the renewal equation (5.15) specializes to

$$(5.18) \quad R_F(t) = r_F(t) + \int_0^t R_F(t - \theta) dF(\theta), \quad t \geq 0.$$

Since the mapping r_F is monotone nonincreasing and takes nonnegative values, it is therefore integrable as a result of (5.11), whence directly Riemann integrable [13, pp. 190-191]. The fact that $0 \leq r_F(t) \leq m_F(r)$ for all $t \geq 0$ implies that R_F is bounded on finite intervals [13, Thm. 4.2, p. 184]. Finally, note that the distribution F has support on \mathbb{N} and is therefore arithmetic, say with span d . All requisite conditions are now in place to apply the basic renewal theorem [13, Thm. 5.5.1, p. 191] to the renewal equation (5.18) to obtain

$$(5.19) \quad \lim_n R_F(c + nd) = \frac{d}{m_F} \sum_{n=0}^{\infty} r_F(c + nd), \quad c \geq 0$$

where m_F is the first moment of F (which is finite by Proposition 5.3). Since the mapping r_F is nonincreasing, it readily follows from (5.19) that for all $c \geq 0$,

$$(5.20) \quad \begin{aligned} \lim_n R_F(c + nd) &\leq \frac{1}{m_F} \left\{ dr_F(0) + \sum_{n=1}^{\infty} dr_F(nd) \right\} \\ &\leq \frac{1}{m_F} \left\{ dr_F(0) + \int_0^{\infty} r_F(t) dt \right\} \\ &= \frac{1}{m_F} \{dm_F(r) + K_F(r)\} < \infty, \end{aligned}$$

where the finiteness of the last bound results from (5.10)–(5.11). In particular,

$$(5.21) \quad \lim_n R_F(nd + \ell) \leq \frac{1}{m_F} \{dm_F(r) + K_F(r)\} < \infty, \quad \ell = 1, 2, \dots, d;$$

and therefore

$$(5.22) \quad \sup_n R_F(n) < \infty.$$

Since $N(t)$ is constant on $[n, n+1)$, direct inspection of (5.9) shows that $R_F(t) \leq R_F(n)$ whenever $n \leq t < n+1$ owing to the monotonicity of r , whence $\sup_{t \geq 0} R_F(t) = \sup_n R_F(n)$ and (5.17) is now immediate from (5.22). \square

Proof of Theorem 5.1. With r given by $r(x) = x^{\gamma-1}$ for all $x \geq 0$, observe that

$$(5.23) \quad K_F(r) = \int_0^{\infty} \int_0^{\theta} r(\theta - t) dt dF(\theta) = \frac{1}{\gamma} \int_0^{\infty} \theta^{\gamma} dF(\theta).$$

Under **(R4)**, Proposition 5.3 now implies the conditions (5.10)–(5.11), and a straightforward application of Theorem 5.4 yields (5.1). \square

5.4. Extensions. Bounds of the form (5.1) are related to the stability of the system under the class of policies of interest and are typically established through system-specific arguments. In fact, as will become apparent from the discussion in §§6 and 7, (5.1) need only hold for a small number of policies. The methods of the present section can be extended to the general model in the following way. Let Π denote a class of policies under which (5.1) is sought to hold, and consider the conditions **(G1)**–**(G3)** below, where

(G1) There exists a positive constant c_s such that for every policy π in Π ,

$$(5.24) \quad P^{\pi} [|X(n+1)| \leq |X(n)| - c_s] = 0, \quad n = 1, 2, \dots$$

Define the rvs $\{\tau_k, \sigma_k, \theta_k, k = 1, 2, \dots\}$ and $\{N(t), t \geq 0\}$ as in §5.2, and assume all these rvs to be finite almost surely under each policy π in Π . Under **(G1)**, it follows that for every π in Π ,

$$(5.25) \quad |X(n)| \leq c_s(\tau_{N(n)+1} - n) \quad P^\pi - \text{a.s.}, \quad n = 1, 2, \dots$$

In this general context, $\{N(t), t \geq 0\}$ may not be a renewal process, since π need not be a stationary policy. With ℓ a positive integer, introduce conditions **(G2)**, **(G3)** as

(G2) There exists a stationary policy π^* in Π such that either

$$E^\pi [|X(n)|^{\ell-1} | X(0) = x] \leq E^{\pi^*} [|X(n)|^{\ell-1} | X(0) = x],$$

or

$$E^\pi [|\tau_{N(n)+1} - n|^{\ell-1} | X(0) = x] \leq E^{\pi^*} [|\tau_{N(n)+1} - n|^{\ell-1} | X(0) = x]$$

$$x \in \mathbb{N}^K, \quad n = 1, 2, \dots$$

for every policy π in Π ; and

(G3) For every initial condition x in \mathbb{N}^K , there exists a positive constant $C_\ell(x)$ such that

$$E^{\pi^*} [\tau_1^\ell | X(0) = x] \leq C_\ell(x).$$

Conditions **(G1)**–**(G2)** imply via (5.25) that for every policy π in Π ,

$$(5.26) \quad E^\pi [|X(n)|^{\ell-1}] \leq c_s^{\ell-1} E^{\pi^*} [|\tau_{N(n)+1} - n|^{\ell-1}], \quad n = 1, 2, \dots$$

Assumptions **(G1)**–**(G3)** are natural in queueing systems when conservation laws are available. Note that $|\cdot|$ could denote *any* norm on \mathbb{R}^K , a fact that could be used to advantage when employing conservation laws. The generalization of Theorem 5.1 can now be stated.

THEOREM 5.1BIS. *Consider the general model. Under **(G1)**–**(G3)**, there exists a single constant K_γ (with $\gamma = l$) such that (5.1) holds for every policy π in Π .*

Proof. Under the Markov stationary policy π^* , the process $\{N(t), t \geq 0\}$ is a delayed renewal process. The proof of (5.1) with $\pi = \pi^*$ is identical to the proof of Theorem 5.1 and the desired result now follows from **(G2)**. \square

6. On the Poisson equation.

6.1. The Poisson equation. Fix η in the unit interval $[0, 1]$ and denote by P^η (respectively, E^η) the probability measure (respectively, expectation operator) induced by the policy f^η . Moreover, let P_x^η (respectively, E_x^η) denote the (conditional) probability measure (respectively, expectation operator) induced by the policy f^η given that $X(0) = x$, with x ranging in \mathbb{N}^K .

Recall that under P^η , the rvs $\{X(n), n = 0, 1, \dots\}$ form a time-homogeneous Markov chain over \mathbb{N}^K , and let $(P^\eta(x, y))$ denote the corresponding one-step transition probabilities. It is plain from (4.3) that

$$(6.1) \quad P^\eta(x, y) = \eta P^1(x, y) + (1 - \eta) P^0(x, y), \quad x, y \in \mathbb{N}^K$$

where $(P^1(x, y))$ (respectively, $(P^0(x, y))$) are the one-step transition probabilities under \bar{g} (respectively, \underline{g}).

The mapping $h : \mathbb{N}^K \rightarrow \mathbb{R}$ and the scalar J are said to solve the *Poisson equation* (associated with the policy f^η) with forcing function $c : \mathbb{N}^K \rightarrow \mathbb{R}$ if

$$(6.2) \quad h(x) + J = c(x) + \sum_y P^\eta(x, y)h(y), \quad x \in \mathbb{N}^K.$$

Clearly the solution pair (h, J) to (6.2) depends on η (and on c), and it is the purpose of this section to establish its regularity properties with respect to η . This information is essential both for establishing the validity of the certainty equivalence principle [22], [30] and for studying the convergence of the stochastic approximation algorithm (4.6) by the method of Metivier and Priouret [23]. From now on, this dependence of J and $h(x)$ on the bias η is denoted simply by $J_c(\eta)$ and $h(\eta, x)$ for all x in \mathbb{N}^K .

Define the *first return time* to state $x = 0$ as the \mathcal{F}_n -stopping time T given by

$$(6.3) \quad T := \inf\{n > 0 : X(n) = 0\}$$

so that $T = \tau_1$ in the notation of §5. Set

$$(6.4) \quad T_\ell(x) := \mathcal{E}_x[T^\ell] = E_x^\eta[T^\ell], \quad x \in \mathbb{N}^K \quad \ell = 1, \dots, \gamma,$$

where the notation that follows Proposition 5.3 has been used. For easy reference recall the estimate (5.5), valid under **(R1)**–**(R4)**, i.e., for each $\ell = 1, \dots, \gamma$, there exists a positive constant C_ℓ so that

$$(6.5) \quad T_\ell(x) \leq C_\ell(1 + |x|^\ell), \quad x \in \mathbb{N}^K.$$

As pointed out already in Section 5, during each slot, at most one customer may leave the system, so that for each $t = 0, 1, \dots$, $|X(t)|$ is necessarily no larger than the forward recurrence time (expressed in slots) to the empty state, and in particular $|X(0)| \leq T$. Since the mapping $x \rightarrow \tilde{c}(|x|)$ defined in **(R5)** is a nondecreasing function of $|x|$, it is plain from (6.5) that whenever $\delta + 1 \leq \gamma$, the bounds

$$(6.6) \quad E_x^\eta \left[\sum_{t=0}^{T-1} |c(X(t))| \right] \leq E_x^\eta \left[\sum_{t=0}^{T-1} \tilde{c}(|X(t)|) \right] \leq \mathcal{E}_x [T\tilde{c}(T)] < \infty, \quad x \in \mathbb{N}^K$$

hold, and the definition

$$(6.7) \quad C(\eta, x) := E_x^\eta \left[\sum_{i=0}^{T-1} c(X(i)) \right], \quad x \in \mathbb{N}^K$$

is thus well posed. An explicit expression for a solution to the Poisson equation is available and is now given [9], [30].

THEOREM 6.1. *Assume conditions **(R1)**–**(R6)** to hold with $\rho < 1$ and $\delta + 1 \leq \gamma$. A solution pair $(h(\eta), J_c(\eta))$ to the Poisson equation (6.2) with $h(\eta, 0) = 0$ is given by*

$$(6.8a) \quad J_c(\eta) = \frac{C(\eta, 0)}{T_1(0)} \quad \text{and} \quad h(\eta, x) = C(\eta, x) - J_c(\eta)T_1(x), \quad x \in \mathbb{N}^K$$

and the equality

$$(6.8b) \quad J_c(f^\eta) = \lim_n E^\eta \left[\frac{1}{n+1} \sum_{t=0}^n c(X(t)) \right] = J_c(\eta)$$

holds true.

In view of (6.8b) and of the ergodic properties of this system under f^η , it is plain that $J_c(\eta)$ is also the expectation of $c(X)$ under the invariant measure corresponding to the policy f^η .

6.2. Lipschitz continuity. The representation (6.8) will be put to use in studying the regularity of the solution pair to the Poisson equation (6.2). To simplify the presentation of the main result of this section, set

$$(6.9) \quad K(x) := \mathcal{E}_x[T^2\tilde{c}(T)], \quad x \in \mathbf{N}^K.$$

THEOREM 6.2. *Assume (R1)–(R6) with $\rho < 1$ and $\delta + 2 \leq \gamma$. Then for all x in \mathbf{N}^K , $K(x) < \infty$ and the function $\eta \rightarrow C(\eta, x)$ is Lipschitz continuous on $[0, 1]$ with Lipschitz constant $4K(x)$, i.e.,*

$$(6.10) \quad |C(\eta, x) - C(\eta', x)| \leq 4K(x)|\eta - \eta'|, \quad \eta, \eta' \in [0, 1].$$

Proof. Fix x in \mathbf{N}^K . That $K(x)$ and $\mathcal{E}_x[T\tilde{c}(T)]$ are both finite is plain from (6.5) under the assumption $\delta + 2 \leq \gamma$. The result (6.10) is established below for c nonnegative in the form

$$(6.11) \quad |C(\eta, x) - C(\eta', x)| \leq 2K(x)|\eta - \eta'|, \quad \eta, \eta' \in [0, 1]$$

so that the result for a general c is now immediate. Therefore, it suffices to assume c to be nonnegative in the remainder of this proof. The arguments proceed in three steps.

Step 1. Fix η in $[0, 1]$. Note that for every \mathbf{N}^K -valued sequence $\{x(i), i = 0, 1, \dots\}$ with $x(0) = x$, the relations

$$(6.12) \quad P_x^\eta[X(i) = x(i), 1 \leq i \leq m] = \prod_{i=0}^{m-1} P^\eta(x(i), x(i+1)), \quad m = 1, 2, \dots$$

hold as a result of the Markov property of the chain $\{X(n), n = 0, 1, \dots\}$ under P^η . The product form of (6.12) and the linear structure of (6.1) now imply that for each $m = 1, 2, \dots$, the mapping $\eta \rightarrow P_x^\eta[X(i) = x(i), 1 \leq i \leq m]$ is a polynomial of degree m in η over $[0, 1]$ and has derivatives of all orders.

Set $A = [X(i) = x(i), 1 \leq i \leq m]$ in (6.12) and observe that

$$(6.13) \quad \frac{d}{d\eta} P_x^\eta[A] = \sum_{t=0}^{m-1} \left[[P^1(x(t), x(t+1)) - P^0(x(t), x(t+1))] \prod_{i=0, i \neq t}^{m-1} P^\eta(x(i), x(i+1)) \right].$$

This suggests defining for every $t = 0, 1, \dots$, the policy 0_t (respectively, 1_t) as the Markov policy that operates according to f^0 (respectively, f^1) at time t , and according to f^η otherwise. With this notation, (6.13) now takes the form

$$(6.14) \quad \frac{d}{d\eta} P_x^\eta[X(i) = x(i), 1 \leq i \leq m] = \sum_{t=0}^{m-1} [P_x^{1_t}[A] - P_x^{0_t}[A]].$$

The definition of the policies 0_t and 1_t implies that $P_x^{1_t}[A] = P_x^{0_t}[A]$ whenever $m \leq t$, so that (6.14) can also be rewritten as

$$(6.15) \quad \frac{d}{d\eta} P_x^\eta[X(i) = x(i), 1 \leq i \leq m] = \sum_{t=0}^n [P_x^{1_t}[A] - P_x^{0_t}[A]], \quad 1 \leq m \leq n.$$

Step 2. To proceed, define

$$(6.16) \quad C_m(\eta, x) := E_x^\eta \left[I[T \leq m] \sum_{t=0}^{T \wedge m-1} c(X(t)) \right] = E_x^\eta \left[\sum_{k=1}^m I[T = k] \sum_{t=0}^{k-1} c(X(t)) \right]$$

for all $m = 1, 2, \dots$. The definition of T implies that

$$(6.17) \quad [T = k] = [X(t) \neq 0, 0 < t < k, X(k) = 0], \quad k = 1, 2, \dots$$

so that

$$(6.18) \quad C_m(\eta, x) = \sum_{k=1}^m \sum_{(x(1), \dots, x(k)) \in \mathcal{Z}_k} P_x^\eta [X(i) = x(i), 1 \leq i \leq k] \sum_{t=0}^{k-1} c(x(t)),$$

where the second sum is taken over the set \mathcal{Z}_k given by

$$(6.19) \quad \mathcal{Z}_k := \{(x(1), x(2), \dots, x(k)) \in (\mathbb{N}^K)^k : x(i) \neq 0, 1 \leq i < k \text{ and } x(k) = 0\}.$$

$k = 1, 2, \dots$

By arguments made earlier, it is plain that on the event $[T = k]$, the bounds $|X(t)| \leq k$, $0 \leq t \leq k$, must necessarily hold, and therefore (6.18) reduces to

$$(6.20) \quad C_m(\eta, x) = \sum_{k=1}^m \sum_{(x(1), \dots, x(k)) \in \mathcal{Z}'_k} P_x^\eta [X(i) = x(i), 1 \leq i \leq k] \sum_{t=0}^{k-1} c(x(t))$$

where the *finite* set \mathcal{Z}'_k is given by

$$(6.21) \quad \mathcal{Z}'_k := \{(x(1), x(2), \dots, x(k)) \in \mathcal{Z}_k : |x(i)| \leq k, 1 \leq i \leq k\}, \quad k = 1, 2, \dots$$

Hence, in view of remarks made earlier in the proof, the mapping $\eta \rightarrow C_m(\eta, x)$ is a polynomial of degree m in η since it is the sum of a finite number of polynomial functions, each one of degree no greater than m .

Since $C_m(\eta, x)$ is a polynomial in η for each $m = 1, 2, \dots$, the derivative $\dot{C}_m(\eta, x)$ exists in the interval $[0, 1]$. To compute it, differentiate (6.20) and use (6.14)–(6.15) to conclude that

$$(6.22) \quad \dot{C}_m(\eta, x) = \sum_{t=0}^{m-1} \left(E_x^{1_t} \left[\sum_{s=0}^{T \wedge m-1} I[T \leq m] c(X(s)) \right] - E_x^{0_t} \left[\sum_{s=0}^{T \wedge m-1} I[T \leq m] c(X(s)) \right] \right).$$

The very same argument that lead from (6.14) to (6.15) now implies

$$(6.23) \quad E_x^{1_t} \left[I[T = k] \sum_{s=0}^{k-1} c(X(s)) \right] = E_x^{0_t} \left[I[T = k] \sum_{s=0}^{k-1} c(X(s)) \right], \quad 0 \leq k \leq t;$$

and therefore (6.22) can be rewritten (in the manner of (6.16)) as

$$(6.24) \quad \dot{C}_m(\eta, x) = \sum_{t=0}^{m-1} \left(E_x^{1_t} \left[\sum_{s=0}^{T \wedge m-1} I[t < T \leq m] c(X(s)) \right] - E_x^{0_t} \left[\sum_{s=0}^{T \wedge m-1} I[t < T \leq m] c(X(s)) \right] \right).$$

On the other hand,

$$\begin{aligned}
 \left| \sum_{t=0}^{m-1} E_x^{1_t} \left[\sum_{s=0}^{T \wedge m-1} I[t < T \leq m] c(X(s)) \right] \right| &\leq \sum_{t=0}^{m-1} \left| E_x^{1_t} \left[I[t < T] \sum_{s=0}^{T \wedge m-1} I[T \leq m] \tilde{c}(T) \right] \right| \\
 &\leq \sum_{t=0}^{m-1} \mathcal{E}_x [I[t < T \leq m] T \tilde{c}(T)] \\
 (6.25) \qquad \qquad \qquad &\leq \mathcal{E}_x [T^2 \tilde{c}(T)]
 \end{aligned}$$

by elementary calculations. A similar bound holds for the terms corresponding to the policies 0_t in (6.24). It then follows from (6.24) and (6.25) that the derivative $\dot{C}_m(\eta, x)$ of $C_m(\eta, x)$ is bounded on $[0, 1]$ by $2K(x)$, and this uniformly in m , i.e.,

$$(6.26) \qquad \qquad \left| \dot{C}_m(\eta, x) \right| \leq 2K(x), \quad \eta \in [0, 1], \quad m = 0, 1, \dots$$

Step 3. The easy estimates

$$\begin{aligned}
 (6.27) \qquad \qquad 0 \leq C(\eta, x) - C_m(\eta, x) &= E_x^\eta [I[T > m] \sum_{t=0}^{T-1} c(X(t))] \leq E_x^\eta [I[T > m] T \tilde{c}(T)], \\
 &\qquad \qquad \qquad m = 0, 1, \dots
 \end{aligned}$$

imply via the monotone convergence theorem that $\lim_m C_m(\eta, x) = C(\eta, x)$ uniformly in η since $\mathcal{E}_x[T \tilde{c}(T)] < \infty$. Consequently, with $0 \leq \eta < \eta' \leq 1$,

$$\begin{aligned}
 |C(\eta, x) - C(\eta', x)| &= \lim_m |C_m(\eta, x) - C_m(\eta', x)| \\
 &= \lim_m \left| \int_\eta^{\eta'} \dot{C}_m(y, x) dy \right| \\
 (6.28) \qquad \qquad &\leq 2K(x) |\eta - \eta'|
 \end{aligned}$$

upon making use of (6.26), and this establishes (6.11). \square

Note that the estimate (6.27) shows that $C(\eta, x)$ is *continuous* under the weaker condition $\delta + 1 \leq \gamma$.

6.3. Corollaries. Theorem 6.2 has several useful consequences that are given in the next few corollaries. The first such corollary is obtained by combining Theorems 6.1 and 6.2 in a straightforward manner; details are left to the interested reader.

COROLLARY 6.3. *Under the hypotheses of Theorem 6.2, the functions $\eta \rightarrow J_c(\eta)$ and $\eta \rightarrow h(\eta, x)$, with x ranging in \mathbb{N}^K , are Lipschitz continuous on $[0, 1]$, i.e., for all η and η' in $[0, 1]$,*

$$(6.29) \qquad \qquad |J_c(\eta) - J_c(\eta')| \leq 4 \frac{K(0)}{T_1(0)} \cdot |\eta - \eta'|$$

and

$$(6.30) \qquad \qquad |h(\eta, x) - h(\eta', x)| \leq 4K_h(x) \cdot |\eta - \eta'|, \quad x \in \mathbb{N}^K$$

with

$$(6.31) \quad K_h(x) := K(x) + \frac{K(0)}{T_1(0)} \cdot T_1(x), \quad x \in \mathbb{N}^K.$$

The behavior of the Lipschitz constants $K(x)$ and $K_h(x)$, and of the solution $h(\eta, x)$ for $|x|$ large is needed in some of the arguments given in §7. The estimates on the Lipschitz constants are given first.

COROLLARY 6.4. *Assume **(R1)**–**(R6)** with $\rho < 1$ and $\delta + 2 \leq \gamma$. There exists a positive constant C such that*

$$(6.32a) \quad |K(x)| \leq C(1 + |x|^{\delta+2}), \quad x \in \mathbb{N}^K$$

and

$$(6.32b) \quad |K_h(x)| \leq C(1 + |x|^{\delta+2}), \quad x \in \mathbb{N}^K.$$

Proof. Fix $x \in \mathbb{N}^K$. Note from **(R6)** and (6.9) that

$$(6.33) \quad \begin{aligned} K(x) &\leq L\mathcal{E}_x [T^2(1 + T^\delta)] \\ &\leq 2L\mathcal{E}_x [T^{\delta+2}] \leq 2LC_{\delta+2} (1 + |x|^{\delta+2}) \end{aligned}$$

with the last inequality following from (6.5), so that (6.32a) holds wherever $C \geq 2LC_{\delta+2}$. The inequality (6.32b) is readily obtained from (6.31) upon making use of (6.5) and (6.32a). \square

The growth of solutions to the Poisson equation can now be described.

COROLLARY 6.5. *Assume **(R1)**–**(R6)** with $\rho < 1$ and $\delta + 1 \leq \gamma$. There exists a positive constant B_h such that*

$$(6.34) \quad |h(\eta, x)| \leq B_h(1 + |x|^{\delta+1}), \quad x \in \mathbb{N}^K$$

for every η in $[0, 1]$.

Proof. By the remark following the proof of Theorem 6.2, the mapping $\eta \rightarrow C(\eta, 0)$ is continuous on $[0, 1]$ and therefore bounded there. For each x in \mathbb{N}^K , straightforward arguments show that

$$(6.35) \quad \begin{aligned} |h(\eta, x)| &\leq E_x^\eta \left[\sum_{t=0}^{T-1} |c(X(t))| \right] + \frac{T_1(x)}{T_1(0)} \cdot \sup_{0 \leq \eta \leq 1} |C(\eta, 0)| \\ &\leq \mathcal{E}_x [T\tilde{c}(T)] + B_1 T_1(x) \end{aligned}$$

with

$$(6.36) \quad B_1 := \frac{1}{T_1(0)} \cdot \sup_{0 \leq \eta \leq 1} |C(\eta, 0)|.$$

The passage from (6.35) to (6.34) is validated by the same arguments as the ones given in the proof of Corollary 6.4. \square

Finally, a bound on the moments of the rvs $\{h(\eta(n), X(n+1)), n = 0, 1, \dots\}$ is obtained.

COROLLARY 6.6. *Assume **(R1)**–**(R6)** with $\rho < 1$ and $r(\delta + 1) + 1 \leq \gamma$ for some nonnegative integer r . Then there exists a positive constant H_r such that*

$$(6.37) \quad \sup_n E^\alpha [|h(\eta(n), X(n+1))|^r] \leq H_r.$$

Proof. For every η in $[0, 1]$, Corollary 6.5 immediately implies

$$(6.38) \quad |h(\eta, x)|^r \leq |2B_h|^r (1 + |x|^{r(\delta+1)}), \quad x \in \mathbb{N}^K$$

so that

$$(6.39) \quad E^\alpha [|h(\eta(n), X(n+1))|^r] \leq |2B_h|^r (1 + E^\alpha [|X(n+1)|^{r(\delta+1)}]), \quad n = 0, 1, \dots$$

The conclusion (6.37) now follows from Theorem 5.1 with $H_r = |2B_h|^r (1 + K_\gamma)$ since $r(\delta + 1) \leq \gamma - 1$. \square

6.4. The general model. In [30] the authors have developed a methodology for proving smoothness properties of solutions to the Poisson equation in a fairly general setting. This is done by invoking the recurrence properties of the underlying Markov chain in order to obtain continuity, Lipschitz continuity and differentiability properties. The ideas of the present paper are however amenable to generalization as follows. Suppose (as in (4.3)) that at each step the stationary policy \bar{g} is used with some probability η while the stationary policy g is used with probability $(1 - \eta)$. Then, as in (6.1), the one-step transition probabilities take the form

$$(6.40) \quad p_{xy}(\eta) = \eta p_{xy}(0) + (1 - \eta) p_{xy}(1), \quad x, y \in \mathbb{N}^K, \quad \eta \in U$$

where $(p_{xy}(0))$ (respectively, $(p_{xy}(1))$), are the one-step transition probabilities under \bar{g} (respectively, g). Under these assumptions the original MDP collapses to the model where the action space is $U = [0, 1]$, and the transitions are realized according to (6.40).

However, the structure (6.40) may arise through a mechanism different from the one outlined above. Given this structure, with some abuse of notation, let η also denote the stationary policy which uses action η at every stage. Fix η in $[0, 1]$ and, as in (6.13)–(6.14), let 0_t (respectively, 1_t) denote the policy which uses action 0 (respectively, action 1) at time t , and otherwise uses action η . Condition **(G4)** then takes the form

(G4) The distribution of the rv T under the policies P^{1_t} and P^{0_t} is stochastically monotone in t , i.e., for all increasing functions $r : \mathbb{R}_+ \rightarrow \mathbb{R}$, the mappings $t \rightarrow E^{1_t}[r(T)]$ and $t \rightarrow E^{0_t}[r(T)]$ are monotone.

As the arguments in (6.41) below reveal, a condition much weaker than **(G4)** will suffice. Let Π be the collection of policies $\{1_t, 0_t; t = 1, 2, \dots; 0 \leq \eta \leq 1\}$.

THEOREM 6.2BIS. *Consider the general model and assume conditions **(G1)**, **(G2)**, **(G4)**, and **(G3)** for $l = 1, 2, \dots, \gamma$ to hold (with $\tau_1 := T$ as in §5.4). Then the conclusion of Theorem 6.2 holds.*

Proof. For the sake of simplicity, set $c_s = 1$ in **(G1)** as the extension to the case $c_s \neq 1$ is obvious. Under the conditions **(G1)**–**(G3)**, the proof is almost identical to that of Theorem 6.2, except that **(G1)**, **(G4)**, and the monotonicity of \tilde{c} are used in (6.25) to obtain

$$\begin{aligned}
\left| \sum_{t=0}^{m-1} E_x^{1_t} \left[\sum_{s=0}^{T \wedge m-1} I[t < T \leq m] c(X(s)) \right] \right| &\leq \sum_{t=0}^{m-1} \left| E_x^{1_t} \left[I[t < T] \sum_{s=0}^{T \wedge m-1} I[T \leq m] \tilde{c}(T) \right] \right| \\
&\leq \sum_{t=0}^{m-1} |E_x^{1_t} [I[t < T] T \tilde{c}(T)]| \\
(6.41) \qquad \qquad \qquad &\leq \sup_t E_x^{1_t} [T^2 \tilde{c}(T)] \leq C_l(x), \quad x \in \mathbf{N}^K
\end{aligned}$$

with $l = 2 + \delta$. \square

Corollary 6.3 continues to hold as stated under the hypotheses of Theorem 6.2bis. Furthermore, it is easy to see that the growth estimates of Corollaries 6.4 and 6.5 continue to hold under the assumption that

$$(6.42) \qquad C_l(x) \leq \tilde{C}_l \cdot (1 + |x|^{c_m}), \quad x \in \mathbf{N}^K$$

for some positive constants \tilde{C}_l and c_m .

7. Convergence of the stochastic approximations.

7.1. The ODE method. This section is devoted to proving the convergence of the recursive scheme (4.5)–(4.6) under P^α . The discussion is carried out under the assumption that the mapping $\eta \rightarrow J_c(f^\eta)$ is monotone *increasing*. The following additional assumption **(R7)** is imposed in order to carry out the analysis.

(R7) The equation (4.4) has a *unique* solution η^* .

The continuity of $\eta \rightarrow J_c(f^\eta)$ now implies

$$(7.1) \qquad [J_c(f^\eta) - V](\eta - \eta^*) > 0, \quad \eta \neq \eta^* \in [0, 1].$$

The uniqueness of the solution to (4.4) is tantamount to *local strict* monotonicity and in practice, is often verified by establishing some stronger monotonicity property on $\eta \rightarrow J_c(f^\eta)$ such as **(R7b)** below.

(R7b) The mapping $[0, 1] \rightarrow \mathbb{R} : \eta \rightarrow J_c(f^\eta)$ is *strictly monotone increasing*.

In §8, condition **(R7b)** is shown to hold for a steering problem which arises from a constrained optimization problem.

The proof of Theorem 4.1 given below uses a version of the ODE method that was proposed by Metivier and Priouret in [23]. The arguments combine the deterministic lemma of Kushner and Clark [17] with a probabilistic result based on properties of the Poisson equation (6.2). This key result is given in the next proposition, the proof of which is delayed until the second part of the section. To state the result, consider the rvs $\{Y(n), n = 0, 1, \dots\}$ given by

$$(7.2) \qquad Y(n) := J_c(f^{\eta(n)}) - c(X(n+1)), \quad n = 0, 1, \dots$$

and pose

$$(7.3) \qquad m(n, t) := \max\{k > n : \sum_{i=n}^{k-1} a_i \leq t\}, \quad t > 0, \quad n = 0, 1, \dots$$

THEOREM 7.1. Assume **(R1)–(R6)** with $\rho < 1$ and $2\delta + 3 \leq \gamma$. For each $t > 0$, the convergence

$$(7.4) \quad \lim_n \left(\sup_{n \leq k \leq m(n,t)} \left| \sum_{i=n}^k a_i Y(i) \right| \right) = 0 \quad P^\alpha - a.s.$$

takes place.

Proof of Theorem 4.1. As shown in [17], [23], the convergence (7.4) underlines the P^α -almost sure convergence of the estimates $\{\eta(n), n = 0, 1, \dots\}$ to η^* . The reader is invited to consult these references for a complete exposition of the arguments that are now briefly summarized: Interpolate the estimate sequence $\{\eta(n), n = 0, 1, \dots\}$ by a piecewise linear function $\eta^{(0)} : [0, \infty) \rightarrow \mathbb{R}$ such that $\eta^{(0)}(t_n) = \eta(n)$ at time $t_n = \sum_{i=0}^{n-1} a_i$ for all $n = 0, 1, \dots$ (with $t_0 = 0$). Moreover, define a sequence of left shifts $\{\eta^{(n)}(\cdot), n = 0, 1, \dots\}$, i.e., $\eta^{(n)}(t) = \eta^{(0)}(t - t_n)$ for all $t \geq 0$, in order to bring the “asymptotic part” of $\{\eta(n), n = 0, 1, \dots\}$ back to a neighborhood of the time origin.

Now observe that the recursion (4.6) can be written in the form

$$(7.5) \quad \eta(n+1) = \left[\eta(n) + a_{n+1} [(V - J_c(f^{\eta(n)})) + Y(n)] \right]_0^1, \quad n = 0, 1, \dots$$

and that from any convergent subsequence $\{\eta^{(m)}(\cdot), m = 0, 1, \dots\}$ a further convergent subsequence $\{\eta^{(m_p)}(\cdot), p = 0, 1, \dots\}$ can then be extracted by standard boundedness and equicontinuity arguments. It is then easy to see from Theorem 7.1 that the limit $\eta^*(\cdot)$ along this subsequence, and, for that matter, the limit of *any* convergent subsequence, satisfies the ODE

$$(7.6) \quad \dot{\eta}^*(t) = V - J_c(f^{\eta^*(t)}), \quad t \geq 0, \quad \eta^*(0) \in [0, 1].$$

Owing to (7.1), this ODE is *asymptotically stable* with a *unique* stable point η^* in $[0, 1]$. A simple shifting argument now implies $\eta^*(t) = \eta^*$ for all $t \geq 0$ and this completes the proof. These arguments are standard and are therefore omitted here in the interest of brevity. \square

The remainder of this section is devoted to a proof of (7.4).

7.2. A proof of Theorem 7.1. The Poisson equation (6.2) implies the relations

$$(7.7) \quad E^\eta[h(\eta, X(n+1)) \mid \mathcal{F}_n] = h(\eta, X(n)) + J_c(\eta) - c(X(n)), \quad n = 0, 1, \dots$$

for all $0 \leq \eta \leq 1$. It then follows from (6.8b) and (7.2) that

$$(7.8) \quad \begin{aligned} -Y(n) &= c(X(n+1)) - J_c(\eta(n)) \\ &= h(\eta(n), X(n+1)) - E^{\eta(n)}[h(\eta(n), X(n+2)) \mid \mathcal{F}_{n+1}] \\ &= Z_n^{(1)} + Z_n^{(2)} + Z_n^{(3)}, \quad n = 0, 1, \dots, \end{aligned}$$

where

$$(7.9a) \quad Z_n^{(1)} := h(\eta(n), X(n+1)) - E^{\eta(n)}[h(\eta(n), X(n+1)) \mid \mathcal{F}_n]$$

$$(7.9b) \quad Z_n^{(2)} := E^{\eta(n)}[h(\eta(n), X(n+1)) \mid \mathcal{F}_n] - E^{\eta(n+1)}[h(\eta(n+1), X(n+2)) \mid \mathcal{F}_{n+1}]$$

$$(7.9c) \quad Z_n^{(3)} := E^{\eta(n+1)}[h(\eta(n+1), X(n+2)) \mid \mathcal{F}_{n+1}] - E^{\eta(n)}[h(\eta(n), X(n+2)) \mid \mathcal{F}_{n+1}]$$

for all $n = 0, 1, \dots$. It now suffices to show for all $t > 0$ that

$$(7.10) \quad \lim_n \left(\sup_{n \leq \ell \leq m(n,t)} \left| \sum_{i=n}^{\ell} a_i Z_i^{(k)} \right| \right) = 0 \quad P^\alpha - \text{a.s.}, \quad k = 1, 2, 3.$$

This will be done in three steps. To facilitate the presentation, define the rvs $\{S_n^{(k)}, n = 0, 1, \dots\}$, $k = 1, 2, 3$, by

$$(7.11) \quad S_0^{(k)} = 0, \quad S_{n+1}^{(k)} := \sum_{i=0}^n a_i Z_i^{(k)}, \quad k = 1, 2, 3, \quad n = 0, 1, \dots$$

Step 1. The rvs $\{Z_n^{(1)}, n = 0, 1, \dots\}$ form a $(P^\alpha, \mathcal{F}_n)$ -martingale difference, whence $\{S_n^{(1)}, n = 0, 1, \dots\}$ is a zero-mean $(P^\alpha, \mathcal{F}_n)$ -martingale. Routine calculations show that

$$(7.12) \quad \sup_n E^\alpha[|S_n^{(1)}|^2] = \sup_n E^\alpha \left[\sum_{i=0}^n a_i^2 |Z_i^{(1)}|^2 \right]$$

$$(7.13) \quad \leq \sup_n E^\alpha \left[|h(\eta(n), X(n+1))|^2 \right] \cdot 4 \sum_{i=0}^{\infty} a_i^2$$

$$(7.14) \quad \leq 4H_2 \cdot \sum_{i=0}^{\infty} a_i^2.$$

The passage from (7.13) to (7.14) uses the estimate (6.37) given in Corollary 6.6 (with $r = 2$ since $2\delta + 2 \leq \gamma - 1$). It is plain from (2.2) that the left-hand side of (7.12) is finite, and the $(P^\alpha, \mathcal{F}_n)$ -martingale $\{S_n^{(1)}, n = 0, 1, \dots\}$ is thus uniformly integrable under P^α . By the martingale convergence theorem, the rvs $\{S_n^{(1)}, n = 0, 1, \dots\}$ converge almost surely under P^α (to an almost sure finite limit), in which case they form a Cauchy sequence P^α -almost surely and (7.10) follows for $k = 1$.

Step 2. For $k = 2$, define the rvs $\{K_n, n = 0, 1, \dots\}$ by

$$(7.15) \quad K_n := E^{\eta(n)}[h(\eta(n), X(n+1)) \mid \mathcal{F}_n], \quad n = 0, 1, \dots$$

and set

$$(7.16) \quad B_r = \sup_n E^\alpha[|K_n|^r], \quad r = 1, 2, \dots$$

It is clear from (6.37) (with $r = 1, 2$) and Jensen's inequality that $B_1 \leq H_1 < \infty$ and $B_2 \leq H_2 < \infty$.

With this notation, observe that

$$(7.17) \quad \begin{aligned} S_{\ell+1}^{(2)} - S_n^{(2)} &= \sum_{i=n}^{\ell} a_i Z_i^{(2)} \\ &= a_{n-1} K_n - \sum_{i=n}^{\ell} (a_{i-1} - a_i) K_i - a_\ell K_{\ell+1}, \quad 1 \leq n \leq \ell; \end{aligned}$$

therefore

$$(7.18) \quad |S_{\ell+1}^{(2)} - S_n^{(2)}| \leq a_{n-1} |K_n| + \sum_{i=n}^{\ell} (a_{i-1} - a_i) |K_i| + a_\ell |K_{\ell+1}|, \quad 1 \leq n \leq \ell$$

since $a_n \downarrow 0$. If the rvs $\{S_n, n = 1, 2, \dots\}$ and $\{R_n, n = 0, 1, \dots\}$ are now defined by

$$(7.19) \quad S_{n+1} = \sum_{i=0}^n (a_i - a_{i+1}) |K_{i+1}| \quad \text{and} \quad R_n = \sum_{i=0}^n |a_i|^2 |K_{i+1}|^2, \quad n = 0, 1, \dots,$$

then (7.18) becomes

$$(7.20) \quad |S_{\ell+1}^{(2)} - S_n^{(2)}| \leq a_{n-1} |K_n| + |S_\ell - S_{n-1}| + a_\ell |K_{\ell+1}|, \quad 1 \leq n \leq \ell.$$

The definition (7.19) implies

$$(7.21) \quad E^\alpha[S_{n+1}] \leq B_1 \sum_{i=0}^n (a_i - a_{i+1}) = B_1(a_0 - a_{n+1}) \leq B_1 a_0, \quad n = 0, 1, \dots$$

Since $S_n \leq S_{n+1}$, the limit $S_\infty := \lim_n S_n$ exists; therefore $E^\alpha[S_\infty] \leq B_1 a_0$ by using the monotone convergence theorem on (7.21). Consequently, $S_\infty < \infty$ P^α -almost surely and the rvs $\{S_n, n = 0, 1, \dots\}$ form a Cauchy sequence P^α -almost surely, i.e.,

$$(7.22) \quad \lim_n \sup_{\ell \geq n} |S_\ell - S_{n-1}| = 0 \quad P^\alpha - \text{a.s.}$$

To handle the first and last terms of (7.20), observe that $R_n \leq R_{n+1}$, hence the limit rv $R_\infty := \lim_n R_n$ exists and satisfies

$$(7.23) \quad E^\alpha[R_\infty] \leq B_2 \sum_{i=0}^{\infty} a_i^2 < \infty$$

by virtue of the monotone convergence theorem. Consequently, $\lim_n R_n = R_\infty < \infty$ P^α -almost surely, whence $\lim_n a_{n-1} |K_n| = 0$ P^α -almost surely or, equivalently,

$$(7.24) \quad \lim_n \sup_{\ell \geq n} a_{\ell-1} |K_\ell| = 0 \quad P^\alpha - \text{a.s.}$$

by the Cauchy convergence criterion. Making use of (7.22) and (7.24) readily leads (via (7.20)) to the conclusion (7.10) for $k = 2$.

Step 3. For $k = 3$, observe that (7.7) and the estimates of Corollary 6.3 readily yield the estimates

$$(7.25) \quad \begin{aligned} & |E^\eta[h(\eta, X(n+1)) | \mathcal{F}_n] - E^{\tilde{\eta}}[h(\tilde{\eta}, X(n+1)) | \mathcal{F}_n]| \\ &= |h(\eta, X(n)) - h(\tilde{\eta}, X(n)) + J_c(\eta) - J_c(\tilde{\eta})| \\ &\leq 4\tilde{K}(X(n)) \cdot |\eta - \tilde{\eta}|, \quad n = 0, 1, \dots \end{aligned}$$

for all η and $\tilde{\eta}$ in $[0, 1]$, where

$$(7.26) \quad \tilde{K}(x) := K(x) + 2 \frac{K(0)}{T_1(0)} T_1(x), \quad x \in \mathbb{N}.$$

The recursion (4.6) implies

$$(7.27) \quad |\eta(n+1) - \eta(n)| \leq a_{n+1} |V - c(X(n+1))|, \quad n = 0, 1, \dots$$

and the inequality

$$(7.28) \quad |Z_n^{(3)}| \leq 4a_{n+1} Q(X(n+1)), \quad n = 0, 1, \dots$$

is now obtained from (7.25), upon setting

$$(7.29) \quad Q(x) := \tilde{K}(x)(V + |c(x)|), \quad x \in \mathbb{N}^K.$$

Under **(R5)**, with the help of (6.5) and (6.32a), it is a simple exercise to check that

$$(7.30) \quad Q(x) \leq C(1 + |x|^{2\delta+2}), \quad x \in \mathbb{N}^K$$

for some positive constant C . Consequently,

$$(7.31) \quad \begin{aligned} E^\alpha \left[\sum_{i=0}^n a_i |Z_i^{(3)}| \right] &\leq C \cdot \sum_{i=0}^n a_i^2 E^\alpha \left[1 + |X(i+1)|^{2\delta+2} \right] \\ &\leq C(1 + K_\gamma) \cdot \sum_{i=0}^\infty a_i^2, \quad n = 0, 1, \dots, \end{aligned}$$

where the last inequality is a simple consequence of (5.1) (since $2\delta + 2 \leq \gamma - 1$). Now, in exactly the same way as in Step 2 of the proof, the uniform bound (7.31) implies

$$(7.32) \quad \lim_n \sup_{\ell \geq n} \left(\sum_{i=n}^\ell a_i |Z_i^{(3)}| \right) = 0 \quad P^\alpha - \text{a.s.},$$

and (7.9) obviously holds for $k = 3$. \square

7.3. The general model. The results of this section rely on boundedness and smoothness properties of solutions to the Poisson equation, but the structure of the proof is otherwise quite general. In fact, consider a set of stationary policies, parameterized by η in $[0, 1]$ and let $(J_c(\eta), h(\eta, x))$ denote the solution to the Poisson equation under policy η . Such parameterization may arise as in §6.4, but for the purposes of the present section this is immaterial. Suppose that the properties **(i)**–**(iii)** below can be established, as was done for the queueing model under consideration, where

(i) For each η in $[0, 1]$, $J_c(\eta)$ equals the cost under η ,

(ii) For each η in $[0, 1]$, $x \rightarrow h(\eta, x)$ is at most polynomial in x ,

(iii) For each x in \mathbb{N}^K , $\eta \rightarrow h(\eta, x)$ and $\eta \rightarrow J_c(\eta)$ are Lipschitz in η , where the Lipschitz constant of $\eta \rightarrow h(\eta, x)$ is at most polynomial in x .

Then the conclusions of Theorem 7.1 (and hence the conclusions of Theorem 4.1) hold with proofs almost unchanged, provided appropriate bounds on moments of the state process are available. Condition **(ii)** is obtained in Corollary 6.6 and its generalizations, and validates Steps 1 and 2 in §7.2. Conditions **(iii)** allows the argument in Step 3 to be carried through and the only changes required involve the constants and the exponents δ , γ and r .

8. Convergence of the adaptive policy and applications. This final section contains a proof of Theorem 4.2, as well as the discussion of an application that arises in constrained optimization.

8.1. A proof of Theorem 4.2. The proof follows from general results obtained by the authors on the certainty equivalence principle when specialized to “simply randomized” policies [30]. First note that the (assumed) convergence $\lim_n \eta(n) = \eta^*$ in probability under P^α , when combined with Theorem 7.2 of [30], implies the key convergence condition **(C)** [30, §4]. Consequently, the convergence (4.11)–(4.12) follows

from Theorem 3.1bis in [30] provided the hypotheses of Theorems 4.2 and 6.1bis of [30] are satisfied. These hypotheses consist in the tightness of the rvs $\{X(n), n = 0, 1, \dots\}$ under P^α and of bounds on the moments of the rvs $\{c(X(n)), h(\eta^*, X(n)), n = 0, 1, \dots\}$ under various policies. It is easy to check that these conditions are all implied by the following conditions.

There exist $\epsilon > 0$ and a positive constant C_ϵ such that for every nonidling policy π in \mathcal{P} , the bounds

$$(8.1) \quad \sup_n E^\pi [|X(n)|^{1+\epsilon}] \leq C_\epsilon,$$

$$(8.2) \quad \sup_n E^\pi [|c(X(n))|^{1+\epsilon}] \leq C_\epsilon$$

and

$$(8.3) \quad \sup_n E^\pi [|h(\eta^*, X(n))|^{1+\epsilon}] \leq C_\epsilon$$

hold.

By virtue of Theorem 5.1, the bound (8.1) readily holds whenever $1 + \epsilon \leq \gamma - 1$. By assumption, c is of polynomial growth with rate δ , so that (8.2) holds if $\delta(1 + \epsilon) \leq \gamma - 1$ by the remark made earlier. To obtain the third bound (8.3), observe from (6.34) that for every $\epsilon > 0$,

$$(8.4) \quad |h(\eta^*, X(n))|^{1+\epsilon} \leq |2B_h|^{1+\epsilon} (1 + |X(n)|^{(\delta+1)(\epsilon+1)}), \quad n = 0, 1, \dots$$

and (8.3) follows with $(1 + \epsilon)(1 + \delta) \leq \gamma - 1$ by again making use of Theorem 5.1. Consequently (8.1)–(8.3) will hold provided ϵ is chosen positive such that $1 + (1 + \delta)(1 + \epsilon) \leq \gamma$.

An identical analysis applies for the long-run average cost associated with d ; details are left to the interested reader. \square

8.2. An application to constrained optimization. Consider the following situation discussed by Nain and Ross in [24]. Several types of traffic, say voice, video, and data, compete for the use of a single resource (or server). The performance requirements are defined by the minimization of a weighted average of the number of video and data packets subject to the constraint that the average number of voice packets waiting for service does not exceed V . This situation can be modeled by a system of K competing queues with $P = 0$. For a precise definition of the performance measures, set

$$(8.5) \quad c(x) := x_K \quad \text{and} \quad d(x) := \sum_{k=1}^{K-1} d_k x_k, \quad x \in \mathbb{N}^K$$

for positive constants d_1, \dots, d_{K-1} , and denote by $J_c(\pi)$ (respectively, $J_d(\pi)$) the long-run average cost (4.1) associated with the cost c (respectively, d) when using the policy π in \mathcal{P} . The constrained optimization problem (\mathbf{P}_V) is then formulated as

$$(8.6) \quad (\mathbf{P}_V) \quad \text{Minimize } J_d(\cdot) \text{ over } \mathcal{P}_V,$$

where $\mathcal{P}_V := \{\pi \in \mathcal{P} : J_c(\pi) \leq V\}$.

Assume the problem to be feasible and nontrivial, i.e., \mathcal{P}_V is nonempty and the policies which minimize J_d are not in \mathcal{P}_V . In that case, Nain and Ross [24] showed that there exist two strict priority policies \bar{g} and \underline{g} and a bias η^* satisfying the equation

$$(8.7) \quad J_c(f^\eta) = V, \quad \eta \in [0, 1]$$

such that f^{η^*} defined through (4.3) is optimal. While the policies \bar{g} and \underline{g} can be found explicitly, the determination of η^* is a difficult task since for $0 < \eta < 1$ the evaluation of $J_c(f^\eta)$ requires solving a Riemann–Hilbert problem. That this computational difficulty can be circumvented by making use of a stochastic approximation-based policy is the content of the following theorem.

THEOREM 8.1. *Assume (R1)–(R5) with $\rho < 1$ and $\gamma \geq 5$. The scheme (4.5), (4.6) solves the constrained optimization problem (\mathbf{P}_V) provided it is feasible.*

Proof. As shown by Nain and Ross [24, Thm. 3.1, pp. 885–886], if the problem is feasible and nontrivial, then there exist Markov stationary policies \bar{g} and \underline{g} such that (8.7) has at least one solution. In fact, both policies are fixed priority policies with \underline{g} giving highest priority to queue K , and \bar{g} giving lowest priority to queue K , while the relative priorities of the other queues are otherwise identical. Moreover, the mapping $\eta \rightarrow J_d(f^\eta)$ is monotone nondecreasing, in fact strictly monotone increasing as shown in Lemma 8.2 below. When $\gamma \geq 5$, the conditions of Theorems 4.1 and 4.2 are readily verified with $\delta = 1$. Hence, $\lim_n \eta(n) = \eta^*$ P^α -almost surely so that $J_c(\alpha) = J_c(f^{\eta^*}) = V$ and $J_d(\alpha) = J_d(f^{\eta^*})$, i.e., α is a policy in \mathcal{P}_V and is thus also constrained optimal.

If the problem is trivial, i.e., $J_c(\bar{g}) \leq V$, then \bar{g} solves (\mathbf{P}_V) . In that case, the same arguments imply that $\lim_n \eta(n) = 1$ P^α -almost surely and optimality follows. \square

In the case $K = 2$, the two policies \bar{g} and \underline{g} are necessarily the fixed priority rules for queue 1 and 2, respectively. In this case, the adaptive policy does not assume any prior information on the statistics of the system, provided (R1)–(R5) hold with $\gamma \geq 5$. In this case, the optimality was obtained by Shwartz and Makowski [28] under a slightly weaker assumption (namely $\gamma \geq 3$), but the convergence (4.10) was only in probability.

This section concludes with the following monotonicity result which was needed in the proof of Theorem 8.1.

LEMMA 8.2. *Under (R1)–(R5) the mapping $\eta \rightarrow J_c(f^\eta)$ is strictly monotone increasing on $[0, 1]$.*

Proof. It is plain from (6.8) that proving the strict monotonicity of $\eta \rightarrow J_c(f^\eta)$ is equivalent to proving the same for $\eta \rightarrow C(\eta, 0)$. Fix η in $[0, 1]$ and recall the definition (4.3) of the policy f^η .

The representation (6.22) of the derivative of $C_m(\eta, 0)$ can be written in the form

$$(8.8) \quad \begin{aligned} \dot{C}_m(\eta, 0) &= \sum_{t=0}^{m-1} E_0^{1t} \left[\sum_{\ell=1}^m I[T = \ell] \sum_{s=0}^{\ell-1} I[T \leq m] X_K(s) \right] - E_0^{0t} \left[\sum_{\ell=1}^m I[T = \ell] \sum_{s=0}^{\ell-1} I[T \leq m] X_K(s) \right] \\ &= \sum_{t=0}^{m-1} \sum_{\ell=t+1}^m \left[E_0^{1t} \left[I[T = \ell] \sum_{s=0}^{\ell-1} X_K(s) \right] - E_0^{0t} \left[I[T = \ell] \sum_{s=0}^{\ell-1} X_K(s) \right] \right], \end{aligned}$$

where (6.23) was used. If it were possible to show bounds of the form

$$(8.9) \quad \Delta(\ell, t, s) := E_0^{1t} [I[T = \ell] X_K(s)] - E_0^{0t} [I[T = \ell] X_K(s)] \geq \epsilon(\ell, t, s)$$

with $\epsilon(\ell, t, s) \geq 0$ for all $0 \leq s < \ell$ and $0 \leq t < \ell$, and $\epsilon(\ell, t, s) > 0$ for at least one such triple (ℓ, t, s) , then necessarily for *some* m , $0 < \dot{C}_m(\eta, 0) \leq \dot{C}_{m+1}(\eta, 0) \leq \dots$ and the strict monotonicity would follow from the second equality in (6.28).

Fix t and ℓ such that $0 \leq t < \ell$. It is easy to see that $\Delta(\ell, t, s) = 0$ whenever $0 \leq s \leq t < \ell$, so that only the case $0 < t < s$ has to be considered to prove (8.9). This is done by the following coupling arguments.

Let \hat{P} be a probability measure on (Ω, \mathcal{F}) under which **(P1)**–**(P3)** hold and $X(0) = 0$. Moreover, let $\{\beta(n), n = 0, 1, \dots\}$ be a sequence of i.i.d. Bernoulli rvs with parameter η which is also independent of the rvs $\{A(n), B(n), n = 0, 1, \dots\}$ under \hat{P} .

The key point of the proof is to construct on Ω a pair of processes $\{X^0(n), n = 0, 1, \dots\}$ and $\{X^1(n), n = 0, 1, \dots\}$ such that (i) $\{X^0(n), n = 0, 1, \dots\}$ (respectively, $\{X^1(n), n = 0, 1, \dots\}$) under \hat{P} is statistically indistinguishable from $\{X(n), n = 0, 1, \dots\}$ under P_0^{0t} (respectively, P_0^{1t}), and (ii) a simple comparison leads to (8.9). To that end, for each $i = 0, 1$, define the process $\{X^i(n), n = 0, 1, \dots\}$ by the recursion

$$(8.10) \quad \begin{aligned} X_k^i(0) &= 0, & X_k^i(n+1) &= X_k^i(n) + A_k(n) - I[X_k^i(n) \neq 0]U_k^i(n)B_k^i(n) \\ & & k &= 1, \dots, K, \quad n = 0, 1, \dots \end{aligned}$$

where the sequences $\{U^i(n), n = 0, 1, \dots\}$ and $\{B^i(n), n = 0, 1, \dots\}$ still need to be specified.

For $i = 0, 1$, the control actions $\{U^i(n), n = 0, 1, \dots\}$ are defined by

$$(8.11a) \quad U^i(n) := \beta(n)\bar{g}(X^i(n)) + (1 - \beta(n))\underline{g}(X^i(n)), \quad n \neq t$$

$$(8.11b) \quad U^0(t) := \underline{g}(X^0(t)) \quad \text{and} \quad U^1(t) := \bar{g}(X^1(t))$$

so that the rvs $\{X^0(n), n = 0, 1, \dots\}$ (respectively, $\{X^1(n), n = 0, 1, \dots\}$) are governed by the policy 0_t (respectively, 1_t).

Only the service sequences $\{B^i(n), n = 0, 1, \dots\}$, $i = 0, 1$, must be specified. First, set $B^0(n) \equiv B(n)$ for all $n = 0, 1, \dots$ and observe from the construction (8.10), (8.11) that the distribution of $\{X^0(n), n = 0, 1, \dots\}$ under \hat{P} obviously coincides with the distribution of $\{X(n), n = 0, 1, \dots\}$ under P_0^{0t} . The construction of the process $\{B^1(n), n = 0, 1, \dots\}$ is somewhat more involved, and is done below. To facilitate the coupling argument, the actual service duration of each customer will be defined in such a way so as to have identical length (for *each* ω in Ω) in both processes. To do this, the rvs $B^1(n)$ are defined in (8.12)–(8.14) so that the number of unsuccessful services experienced by each customer is identical in both systems. Set

$$(8.12) \quad B^1(n) := B(n), \quad n = 0, 1, \dots, t-1$$

and observe from (8.10) that in order to determine the process $\{X^1(n), n = 0, 1, \dots\}$, it suffices to provide the values of $B_k^1(n)$ at times n such that $U^1(n) = e^k$, $k = 1, \dots, K$. For all $i = 0, 1$ and $k = 1, \dots, K$, set

$$(8.13) \quad \tau_k^i(\ell) := \begin{cases} \min\{n \geq 0 & : U^i(n) = e^k\} & \text{if } \ell = 1 \\ \min\{n > \tau_k^i(\ell-1) & : U^i(n) = e^k\}, & \text{if } \ell = 2, 3, \dots \end{cases}$$

and define

$$(8.14) \quad B_k^1(\tau_k^1(\ell)) := B_k(\tau_k^0(\ell)), \quad k = 1, \dots, K, \quad \ell = 1, 2, \dots$$

With these definitions, the actual number of times each customer is served is identical in both systems, while the sequences $\{B(n), n = 0, 1, \dots\}$ (under $P_0^{1\epsilon}$) and $\{B^1(n), n = 0, 1, \dots\}$ (under \hat{P}) are statistically indistinguishable. Consequently, the distribution of $\{X^1(n), n = 0, 1, \dots\}$ under \hat{P} coincides with the distribution of $\{X(n), n = 0, 1, \dots\}$ under $P_0^{1\epsilon}$. Moreover, by construction (with the notation of (6.3)), it is easy to see that $T^0 = T^1$ and $X_K^0(n) \leq X_K^1(n)$ for all $n = 0, 1, \dots$ \hat{P} -almost surely, whence

$$(8.15) \quad \Delta(\ell, t, s) = \hat{E} [I[T^1 = \ell] (X_K^1(s) - X_K^0(s))] \geq 0.$$

Finally, for $s = t + 1$, observe that on the event A given by

$$A := [T^0 = \ell] \cap [X_K^0(t) \neq 0] \cap [X_k^0(t) \neq 0 \text{ for some } k = 1, 2, \dots, K-1] \cap [B_K(t) = 1],$$

the equality $X_K^1(t+1) - X_K^0(t+1) = 1$ holds, and that $\hat{P}[A] > 0$. Consequently,

$$(8.16) \quad \hat{E} [I[T^1 = \ell] [X_K^1(s) - X_K^0(s)]] := \epsilon(\ell, t, t+1) \geq \hat{P}[A] > 0$$

and the result is established. \square

REFERENCES

- [1] E. ALTMAN AND A. SHWARTZ, *Optimal priority assignment with general constraints*, in *Proceedings of the 24th Allerton Conference on Communications, Control and Computing*, Monticello, IL, (1987), pp. 1147–1148.
- [2] ———, *Optimal priority assignment: a time sharing approach*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 1098–1102.
- [3] J. S. BARAS, A. J. DORSEY, AND A. M. MAKOWSKI, *Two competing queues with geometric service requirements and linear costs: the μ -rule is often optimal*, Adv. Appl. Prob., 17 (1985), pp. 186–209.
- [4] J. S. BARAS, D.-J. MA, AND A. M. MAKOWSKI, *K competing queues with geometric service requirements and linear costs: the μ -rule is always optimal*, Systems Control Lett., 6 (1985), pp. 173–180.
- [5] A. BENVENISTE, M. METIVIER, AND P. PRIOURET, *Algorithmes Adaptatifs et Approximations Stochastiques*, Masson, Paris, 1987.
- [6] F. BEUTLER AND K. W. ROSS, *Optimal policies for controlled Markov chains with a constraint*, Math. Anal. Appl., 112 (1985), pp. 236–252.
- [7] V. S. BORKAR, *Controlled Markov chains with constraints*, Sadhana-Indian Academy of Sciences, Journal for Engineering, 15 (1990), pp. 405–413.
- [8] C. BUYUKKOC, P. P. VARAIYA, AND J. WALRAND, *The $c\mu$ -rule revisited*, Adv. Appl. Prob., 17 (1985), pp. 234–235.
- [9] C. DERMAN AND A. F. VEINOTT, JR., *A solution to a countable system of equations arising in Markovian decision processes*, Ann. Math. Stat., 38 (1967), pp. 582–584.
- [10] E. G. GLADYSHEV, *On stochastic approximation*, Theor. Prob. Appl., 10 (1965), pp. 275–278.
- [11] J. M. HARRISON, *A priority queue with discounted linear costs*, Oper. Res., 23 (1975), pp. 270–282.
- [12] O. HERNÁNDEZ-LERMA, *Adaptive Markov Control Processes*, Springer-Verlag, New York, 1989.
- [13] S. KARLIN AND H. TAYLOR, *A First Course in Stochastic Processes*, Academic Press, New York, 1974.
- [14] G.P. KLIMOV, *Time sharing systems*, Theory of Probab. Appl., 19 (1974), pp. 532–553; 23 (1978), pp. 314–321.
- [15] P.R. KUMAR, *A survey of some results in stochastic adaptive control*, SIAM J. Control Optim., 23 (1985), pp. 329–380.
- [16] H. J. KUSHNER, *Approximation and Weak Convergence Methods for Random Processes, with Applications to Stochastic Systems Theory*, MIT Press, Cambridge, 1984.
- [17] H. J. KUSHNER AND D. S. CLARK, *Stochastic Approximation for Constrained and Unconstrained Systems*, Applied Mathematical Sciences, 26, Springer-Verlag, Berlin, 1978.

- [18] D.-J. MA, A. M. MAKOWSKI AND A. SHWARTZ, *Stochastic approximations for finite state Markov chains*, Stochastic Processes and Their Applications, 35 (1990), pp. 27–45.
- [19] A. M. MAKOWSKI AND A. SHWARTZ, *Implementation issues for Markov decision processes*, in Stochastic Differential Systems, Stochastic Control Theory and Applications, W. Fleming and P.-L. Lions, eds., The IMA Volumes in Mathematics and Its Applications, 10, Springer-Verlag, New York (NY), 1988, pp. 323–337.
- [20] ———, *Recurrence properties of a discrete-time single-server network with random routing*, EE Pub., 718, Technion, Haifa, Israel 1989.
- [21] ———, *Analysis and adaptive control of a discrete-time single-server network with random routing*, Technical Report SRC 89-75r1, Systems Research Center, Univ. of Maryland, College Park MD, 1989.
- [22] P. MANDL, *Estimation and control in Markov chains*, Adv. Appl. Prob. 6 (1974), pp. 40–60.
- [23] 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. 140–150.
- [24] P. NAIN AND K. W. ROSS, *Optimal priority assignment with hard constraint*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 883–888.
- [25] S. ROSS, *Introduction to Stochastic Dynamic Programming*, Academic Press, New York, 1983.
- [26] K. W. ROSS, *Constrained Markov Decision Processes with Queueing Applications*, Ph.D. Thesis, Computer, Information and Control Engineering, Univ. of Michigan, Ann Arbor, MI, 1985.
- [27] L.I. SENNOTT, *Constrained average-cost Markov decision chains*, Preprint, 1990.
- [28] 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 in Control and Inform. Sci., 83 (1986). Springer-Verlag, New York, Berlin, pp. 515–532.
- [29] ———, *Comparing policies in Markov decision processes: Mandl's Lemma revisited*, Math. Oper. Res., 15 (1990), pp. 155–174.
- [30] ———, *On the Poisson equation for Markov chains*, Math. Oper. Res., under revision, 1987.
- [31] P. TSOUCAS AND J. WALRAND, *Optimal adaptive server allocation in a network*, Systems Control Lett., 7 (1986), pp. 323–327.